Realistic Object Properties Impact Multisensory Perception & Memory

By

SHEA E. DUARTE
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Psychology

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Joy J. Geng, Chair

Simona Ghetti

Andrew P. Yonelinas

Committee in Charge

2024

To Mom and Dad, who gave me the confidence, curiosity, and unwavering support to pursue this path. And to Josue, for convincing me that I could do it, over and over again.

Acknowledgements

I am endlessly grateful to have had Joy Geng as my advisor. Joy, your mentorship and guidance pushed me to exceed my own expectations of what I could be capable of as a scientist. Our animated discussions about data and brainstorming sessions for new studies have been highlights of my academic journey, and I look forward to more of them in the future. I am also grateful for my dissertation committee, Simona Ghetti and Andrew Yonelinas. Your invaluable time, feedback, and support on grants and papers have been crucial to my growth as a researcher.

I owe a tremendous debt of gratitude to the members of the Geng Lab, who have been both mentors and friends throughout this journey. Special thanks to the postdocs Bo Yeong Won, Joe Nah, Joe Zhou, Martha Forloines, and Jared Stokes, and the graduate students Xinger Yu, Philip Witkowski, Raisa Rahim, and Beth Hall, for the help with code, feedback in lab meetings, and reassurance over the years. Your brilliance and kindness made my time in the lab such a joy.

A huge thank you to the research assistants I have had the privilege to work with, Jocelyn Huerta, Sachi Bhatt, Karan Gupta, Tara Sharifi, Cheska Wang, Zoe Hareng, Eliana Ertsey, Rishit Das, Cailey Tennyson, and Neel Majmudar, thank you for your dedication to these projects and for all of the hours you put into data collection and analysis. Research was always more fun with you all and I am grateful for your enthusiasm and commitment.

To my friends in Davis, Beverly, Forrest, Madison, Toby, Anna, Simran, Lee, Soukhin, and Diego, thank you all for the support, for piloting my experiments, and for counterbalancing this challenge with good times.

And lastly, thank you to my family. Mom, Dad, Gavin, Vovó, Grandma, and so many others. Thank you for everything you did to help me get to where I am, I owe it all to you.

Abstract

Multisensory experiences are ubiquitous in our everyday lives and impact what sensory information we notice, pay attention to, and remember. However, many areas of cognitive psychology focus on the senses individually, and/or use simplistic versions of real objects. This makes it difficult to understand whether, and how, laboratory findings can explain cognitive processing in the real world. This dissertation investigates how naturalistic object properties, including multiple sensory modalities, semantic information, and dynamic motion, contribute to sensory processing and memory formation. Chapter 2 examines the impact of task-irrelevant, semantically congruent sounds on visual recognition memory. Through a series of experiments, it demonstrates that congruent object-sound pairings facilitate recollection-based recognition and promote the formation of multisensory memories. These findings underscore the importance of considering multisensory interactions in developing models of memory applicable to real-world settings. Building on these insights, Chapter 3 investigates how multisensory object processing affects memory for nearby visual objects and scene contexts. While the presence of audiovisual objects at encoding did not significantly benefit memory for nearby visual objects, it did improve recall of the environmental context. These results highlight the broader influence of multisensory processing on episodic memory formation beyond individual objects. Chapter 4 explores the audiovisual ventriloquist effect using realistic stimuli in virtual reality. In this study, we found that animated, semantically congruent audiovisual stimuli show enhanced spatial ventriloquism at small disparities relative to the simplistic stimuli frequently used in laboratory studies of multisensory integration. The study emphasizes the role of stimulus realism and dynamic motion in audiovisual integration. Collectively, this research advances our understanding of how multisensory experiences shape memory and perception in naturalistic settings.

Table of Contents

Chapter 1: Introduction

Our everyday experiences are largely multisensory, that is, we receive sensory information from two or more modalities simultaneously about an event. Imagine walking into a coffee shop and hearing the familiar gurgle of an espresso machine as you spot the stream of dark liquid fill a cup, the bubbling hiss of milk being frothed as a barista plunges the steaming wand in a metal pitcher, and the clatter of a teacup being placed on a saucer. Many early studies of multisensory integration showed that it affects signal processing at a low level by increasing the probability of detecting near-threshold stimuli through additive or even super-additive neuronal firing (Stein et al., 2020). Research on multisensory processing has exploded over recent decades and demonstrated effects that cascade from low-level perception to object perception, attention, memory, and learning (Shams & Seitz, 2008; Macaluso et al., 2016; Matusz et al., 2017; Stein et al., 2020). These studies generally show that multimodal stimuli support object recognition and speech perception, increase attentional capture, and improve aspects of memory and learning. Despite the emergence of such findings and the ubiquity of multimodal stimuli in our environments, most research across the cognitive sciences focuses on single sensory modalities in isolation. Further, the study of multisensory object processing itself has primarily focused on how simple, two-dimensional, transient stimuli are integrated, with studies of realistic stimuli primarily. While these choices have frequently been made in an attempt to maintain tight experimental control and an isolation of confounding variables, new technologies including art software and virtual reality afford researchers the opportunity to study processes with more naturalistic and multisensory stimuli to take a step towards understand how laboratory findings of cognitive processes generalize to the real world. The overarching goal of

the work presented in this dissertation is to extend what has been found in previous studies of memory and multisensory processing to make them increasingly naturalistic.

Multisensory influences on human memory

Chapters 2 and 3 of this dissertation contain several experiments addressing open questions regarding how multisensory processing impacts visual memory. As with many other fields of cognitive psychology, much of what we currently know about human memory comes from studies using stimuli presented within a single modality, such as lists of words or visual objects. However, recent research suggested that audiovisual presentations of objects along with their characteristic sounds can improve later object recognition memory (Lehmann & Murray, 2005; Heikkilä et al., 2015; Thelen et al., 2015; Moran et al., 2013; Matusz et al., 2017). These studies underscore the importance of further understanding the impacts of multisensory processing on the formation of memories. The studies presented in Chapters 2 and 3 extend existing research in this area by investigating how multisensory processing specifically impacts the distinct processes underlying recognition memory and episodic memory formation.

The goal of Chapter 2 is to understand whether audiovisual presentations of objects specifically impact recollection- or familiarity-based recognition memory. While previous studies had shown that recognition memory was improved by multisensory presentation, successful recognition of previously shown objects can be accomplished via two mechanisms that are neurally and behaviorally dissociable: familiarity and recollection (Diana et al., 2007; Eichenbaum et al., 2007; Yonelinas, 2002; Yonelinas et al., 2010, but see Wais et al., 2006). Recollection reflects the retrieval of specific information from an episodic event, such as when or where the event occurred (e.g., recalling from where you know someone you see on the street), whereas familiarity reflects a general measure of memory strength (e.g., knowing you

have seen the person before; Yonelinas, 2002; Yonelinas et al., 2010). The three experiments presented in Chapter 2 therefore use methods derived from the memory literature to assess whether the impacts of multisensory processing on memory impact familiarity, recollection, or both types of recognition memory.

The experiments presented in Chapter 3 build on this work by assessing the effects of multisensory object encoding on memory for neighboring visual objects and background context. It is possible that the improvement of memory for audiovisual objects comes at a cost to memory for other features of the context in which it was encoded if attention is focused on the multimodal object. On the other hand, it is also possible that the presence of audiovisual object supports episodic memory overall, perhaps due to greater elaborative processing on the entirety of the context. To investigate these alternatives, in Chapter 3, we tested the effect of audiovisual object processing on nearby visual objects and source memory for contextual details in which the object was encoded. The results of Chapters 2 and 3 elucidate the unique ways in which multisensory objects are encoded and how this impacts recognition memory and episodic memories for surrounding contextual information. This work bridges research from the memory and multisensory processing literatures to aid in our understanding of memory formation in real-world situations, which are so often multisensory in nature.

Multisensory integration in realistic objects

Chapter 4 addresses the long-standing question within the multisensory processing literature of whether, and how, properties of realistic stimuli impact the tendency to integrate stimuli using a ventriloquist paradigm in virtual reality (VR). The ventriloquist effect occurs when the location of a sound is perceived to come from that of a synchronous visual stimulus (Bruns, 2019). This effect has been exploited by multisensory researchers to investigate how

various stimulus factors affect the tendency of observers to integrate crossmodal stimuli. Most such studies have used simplistic, transient stimuli presented on two-dimensional displays, making it difficult to know how well laboratory findings generalize to real-world crossmodal stimuli (Körding et al., 2007; Rohe & Noppeney, 2015; Slutsky & Recanzone, 2001; Van Wanrooji et al., 2010). In two experiments, we investigate the influence of semantic correspondence between stimuli and dynamic motion on audiovisual integration within a ventriloquist paradigm presented in VR. Further, we employ Bayesian Causal Inference modeling to understand the mechanisms underlying such influences. In this Chapter, we show the benefits of using emerging technology like VR and three-dimensional models of animated, realistic objects to conduct research using stimuli that increasingly resemble those encountered in the real world to answer outstanding research questions.

Together, the findings presented in this dissertation contribute to our knowledge of how perceptual and memory processes are impacted by increasingly naturalistic properties of stimuli. These studies further highlight the utility of combining methods across research fields and taking advantage of emerging methods to increase the generalizability of cognitive research.

Chapter 2: Object memory is multisensory: Task-irrelevant sounds improve recollection

The following chapter consists of a manuscript that has been published in the journal *Psychological Bulletin & Review*

Abstract

Hearing a task-irrelevant sound during object encoding can improve visual recognition memory when the sound is object-congruent (e.g., a dog and a bark). However, previous studies have only used binary old/new memory tests, which do not distinguish between recognition based on the recollection of details about the studied event or stimulus familiarity. In the present research, we hypothesized that hearing a task-irrelevant, but semantically congruent natural sound at encoding would facilitate the formation of richer memory representations, resulting in increased recollection of details of the encoded event. Experiment 1 replicated previous studies showing that participants were more confident about their memory for items that were initially encoded with a congruent sound compared to an incongruent sound. Experiment 2 suggested that congruent object-sound pairings specifically facilitate recollection and not familiarity-based recognition memory, and Experiment 3 demonstrates that this effect was coupled with more accurate memory for audiovisual congruency of the item and sound from encoding rather than other aspect of the episode. These results suggest that even when congruent sounds are task-irrelevant, they promote formation of multisensory memories and subsequent recollection-based retention. Given the ubiquity of encounters with multisensory objects in our everyday lives, considering their impact on episodic memory is integral to building models of memory that apply to naturalistic settings.

Introduction

Multisensory events are ubiquitous in natural environments, and the integration of crossmodal signals has effects that cascade from perception to learning (Stein et al., 2020; Shams & Seitz, 2008). Despite the prevalence and influence of multisensory stimuli, most areas of research in cognition adopt a unisensory perspective, including memory. For example, studies of recognition memory have traditionally used lists of words or objects presented in a single modality. However, recent research has shown that audiovisual presentations of objects along with their characteristic sounds can improve later object recognition memory (Lehmann & Murray, 2005; Heikkilä et al., 2015; Thelen et al., 2015; Moran et al., 2013; Matusz et al., 2017). This underscores the importance of understanding how multisensory perceptual events impact the formation of specific memories.

In an early demonstration of the multisensory advantage in memory, Lehmann & Murray (2005) had participants discriminate between old and new objects that were initially visual-only or presented with an object-congruent or object-incongruent simultaneous sound. Accuracy on "old" trials was higher for objects initially paired with a congruent sound, despite the sounds being completely task-irrelevant. In subsequent studies, multisensory "old" trials were also differentiated via greater BOLD activation in the lateral occipital cortex (Murray et al., 2005), and ERP results showed distinct brain networks involved as early as 60-135 ms post-visual stimulus (Murray et al., 2004). The emerging work in this area highlights the impact of multisensory processing on recognition memory; however, the binary old/new recognition tasks employed in these studies have led to findings that lack specificity as to which memory mechanisms are affected by multisensory presentations. In the present research, we consider how multisensory processing affects two forms of object memory: recollection and familiarity.

Dual-process memory models posit that recognition memory depends on the contribution of two behaviorally and neurally distinguishable processes, namely recollection and familiarity (Yonelinas, 2002; Diana et al., 2007; Yonelinas et al., 2010; but see Wais et al., 2006). Recollection reflects the retrieval of specific information from an episodic event, such as when or where the event occurred (e.g., recalling from where you know someone you see on the street), whereas familiarity reflects a general measure of memory strength (e.g., knowing you have seen the person before) (Yonelinas, 2002; Yonelinas et al., 2010). Research suggests that encoding manipulations can differentially affect recollection or familiarity. One such situation is the *congruency effect*, whereby an encoded noun is better remembered when paired with a semantically congruent adjective (e.g., banana-yellow) than an incongruent adjective (e.g., spinach-ecstatic) (Craik & Tulving, 1975; Atienza et al., 2011; Hashtroudi, 1983). Bein and colleagues (2015) showed a higher proportion of subjective "recollection" responses to items encoded in the semantically congruent condition, coupled with enhanced retrieval of the context word itself, whereas "familiarity" responses did not differ between conditions. Although the Bein et al. (2015) study used pairs of visually encoded words, studies of multisensory effects on object memory using semantically congruent images and natural sounds may similarly yield recollection-specific memory benefits. On the other hand, increasing the perceptual fluency of stimuli during encoding can support both recollection and familiarity (Yonelinas, 2002). Chen & Spence (2010) showed that multisensory processing of congruent image-sound pairs facilitated the identification of visual objects that were perceptually degraded by visual masks. If the memory benefit of multisensory processing is due solely to improved identification and perception of the visual object during study, both recollection and familiarity may be expected to

improve; familiarity may even benefit more than recollection given its relation to priming mechanisms supporting object identification (e.g., Wang & Yonelinas, 2012).

In the present research, we aim to replicate experimental findings demonstrating the benefits of task-irrelevant, congruent sounds on object memory, and delineate whether this effect is driven by improvements to recollection or familiarity-based recognition. We use experimental paradigms derived from the memory literature to address methodological limitations of previous work on multisensory memory. These prior studies have used binary old/new memory tests, from which a single hit rate and false alarm rate are obtained for items in each condition. However, collecting multiple hit and false alarm rates per participant and encoding condition is essential to measure latent memory signals accurately because hit rates alone are susceptible to response biases that obscure the true strength of the underlying memory trace (Brady et al., 2021; Macmillan & Creelman, 1990). For example, a participant might adopt a very stringent criterion and only endorse an item as old if they are very confident and can retrieve many details. We address this limitation by collecting confidence ratings with each old/new recognition response to examine hit and false alarm rates across a range of response criteria (i.e., confidence levels) for each participant and encoding condition.

Our central hypothesis was that hearing a task-irrelevant, but semantically congruent natural sound at encoding would facilitate the formation of a richer memory representation that would support recollection of details of the encoded event. To anticipate our results, in Experiment 1, we replicated findings showing improved recognition memory for visual images of objects originally presented with a congruent sound compared to those presented with an incongruent sound. In Experiment 2, we formally measured both recollection and familiarity-based recognition, and found that congruent sounds during encoding specifically supported

recollection-based recognition. Finally, in Experiment 3, we asked participants to recollect the sound that was associated with each image at encoding (congruent, control, incongruent) and found the highest rates of recollection for objects seen in a congruent audiovisual pair at encoding. Across three experiments using different methods to estimate recollection and familiarity, we found converging evidence that congruent multisensory information during encoding enhances subsequent recollection.

Experiment 1

We first aimed to conceptually replicate previous studies showing generally improved recognition memory for congruent multisensory pairs using a blocked design and a surprise memory task including confidence ratings. Participants completed a within-subjects audiovisual encoding task in which visual items were paired with congruent, incongruent, or meaningless control sounds, followed by a visual-only, surprise recognition test. Importantly, the auditory stimuli had no relevance for the encoding task, which was to determine if the visual object would fit into a standard-size suitcase, and participants were not asked to remember the items. Half of the visual images were overlaid with semi-transparent visual noise during the encoding task that degraded the visibility of the object. This manipulation served to avoid memory ceiling effects (see Heikkila et al., 2015) and to test whether multisensory processing supports memory by improving the perceptual fluency of visually obscured items at encoding. The memory task included four response options to assess whether semantically congruent audiovisual pairs led to higher-confidence recognition memory than incongruent or control pairs, and to calculate hit and false alarm rates for each confidence level to examine the effect of congruency across response criteria.

Method

Participants. Seventy-five students (62 identified as female and 13 identified as male, M_{age} = 19.8 years) from the University of California, Davis, participated in exchange for partial course credit. Nine participants were excluded based on our pre-registered exclusion criteria due to low accuracy (below chance, 50%) on either the encoding task or the recognition memory task. We also administered a debriefing questionnaire, which was used to determine whether participants should be excluded due to a noisy testing environment, exerting little or no effort in completing the study, or a lack of access to consistent audio (due to glitches, volume changes, or a lack of working speakers) (see Supplemental Materials for full list of questions). No participants were excluded in Experiment 1 under these criteria. Our sample size was determined with an a priori power analysis using the python package Pingouin (Vallat, 2018) with power (1-β) set at 0.95 and $\alpha = 0.05$. Prior unpublished data from our laboratory showed an effect of initial sound congruency on recognition memory for visual items with an effect size of $\eta_p^2 = 0.06$, which requires at least 33 participants to detect. To account for poor testing conditions associated with online data collection, we doubled this number to 66, and data were collected until we reached this point post-exclusion. The pre-registered sample size and exclusion criteria can be found on the Open Science Framework (https://osf.io/5uz24/).

Materials. A total of 180 images of three-dimensional (3D) models of common objects (i.e., tools, household objects, vehicles, animals, instruments, recreational equipment, and miscellaneous common items; see Supplemental Materials for a full list of items) were gathered from the Unity Asset Store (https://assetstore.unity.com/3d). Using the Unity Editor, objects were rotated to easy-to-recognize orientations and edited to reflect the position the object typically assumes when making a sound in order to improve the perception of unity between the item and the sound (e.g., the dog model was edited to have an opened mouth, as if it were

barking) (Edmiston & Lupyan, 2015). We used the python package scikit-image (Van der Walt et al., 2014) to remove the image backgrounds, convert them to black and white, and size them to the same dimensions (500 x 500 pixels). The real-world sizes of half of the objects in the images were "small" (small enough to easily fit in a standard suitcase) and the other half were "large." Ninety of these were used in the encoding task, and 90 new objects from the same categories were integrated into the recognition memory task for a total of 180 items. New items in the recognition task were selected from the same categories as the old objects, and because they make similar types of sounds as the old items either on their own (e.g., a rabbit) or when interacted with (e.g., a scooter) (see Supplemental Materials). The old and new items were not counterbalanced across the encoding and recognition tasks but importantly, the old items, which all had associated sounds, were counterbalanced across the six encoding conditions across participants. Thus, while overall recognition discrimination between old and new objects may be different, this would not affect the critical comparison of interest between recognition of items paired with different sounds in the "congruent", "incongruent", or "white noise" encoding conditions. Six different visual masks were manually created using a variety of black, white, and gray geometric shapes arranged in a square the same size as the images. These were used at 100% opacity for the post-stimulus mask, and displayed at 50% opacity when overlaid on top of images as visual noise.

Natural sounds and white noise sounds were obtained from the Multimost Stimulus Set (Schneider et al., 2008) or found on https://findsounds.com/. 90 natural sounds corresponded to the items in the encoding task for the congruent condition, 15 variations of white noise were used for the control condition, and a separate set of 30 natural sounds were used for items in the incongruent condition. The same 30 incongruent sounds were used in every version of the

experiment, and were chosen from the same categories as the visual objects. These were all from the Multimost Stimulus Set, from which all sounds were shown to be identifiable on their own (Schneider et al., 2008). For each version of the experiment, incongruent sounds were randomly paired with visual objects, and these pairs were manually rearranged in cases where the visual object could be expected to make a noise that was at all similar (e.g., the whistle sound would not be paired with the bird image) (see Supplemental Materials for a full list of images, sounds, and combinations used). All sounds were 400 ms in length and amplitude normalized using Audacity (Audacity Team, 2021).

Procedure. Participants completed separate encoding and recognition blocks online via personal computers through the online stimulus presentation software Testable (https://www.testable.org/). Before the encoding task began, a string of sample beeps was played, and participants were asked to adjust their sound level to a comfortable volume and not to alter it for the remainder of the study.

Encoding Block. The encoding block consisted of a size judgement task. Ninety object-sound pairs were presented during this block (30 congruent pairs, 30 control pairs, and 30 incongruent pairs). On each trial, a visual and an auditory stimulus were simultaneously presented for 400 ms, followed by a 600ms post-stimulus mask. The post-stimulus mask functioned to limit continued visual processing of the object in order to accentuate the timing cooccurrence of the visual and auditory stimuli (Kinsbourne & Warrington, 1962). Participants were to respond by clicking an on-screen "yes" button if the visually presented item would fit inside a standard-sized suitcase, and "no" if it would not. Importantly, participants were informed not to pay attention to the sounds, and to base their size judgements on the item in the image. The auditory stimulus was either semantically congruent to the visual stimulus,

incongruent, or a white noise control sound. Additionally, while all presentations were followed by the 600 ms post-stimulus visual mask, half of the items were also overlaid with visual noise during the initial 400 ms presentation. Therefore, there were three levels of Auditory Condition (congruent, incongruent, and control), and two levels of Visual Noise (visual noise, no visual noise) (see Figure 2.1a). Visual stimuli were counterbalanced across the six, within-subjects encoding conditions, and there were 90 trials randomized for each participant. The size judgement task was designed to prevent participants from expecting that their memory might be tested for objects in this block, making the recognition block a test of incidental memory.

Recognition Block. Immediately following the encoding block, participants completed a visual-only surprise recognition task. In this task, the 90 old images were intermixed with 90 new images for a total of 180 trials. On each trial, a visual stimulus was presented for 400 ms, and participants gave a confidence-based recognition response, indicating whether the item was "definitely old," "probably old," "probably new," or "definitely new" (see Figure 2.1b). Trials were randomized for each participant.

Debriefing Questionnaire. After the experiment, participants responded to questions on a debriefing survey, which allowed us to assess the quality of the testing environment and stimulus presentation. The questionnaire included questions about the testing environment, the subjective volume and quality of the auditory stimuli, whether the volume was adjusted during the experiment, whether any glitches or lags between audiovisual stimuli were experienced, among others. As this experiment was completed remotely, responses were used to exclude participants when the testing environment or stimulus presentations were not of adequate quality.

Data Analysis. The design, hypotheses, and statistical analyses for Experiment 1 were preregistered prior to data collection on the Open Science Framework, and raw data files and

analysis code are publicly available (https://osf.io/5uz24/). The preregistered analysis tests for differences in memory performance (indexed by confidence scores) between encoding conditions. We also performed an exploratory receiver operating characteristic (ROC; Yonelinas & Parks, 2007) analysis to assess hit and false alarm rates between encoding conditions at each response criterion. Additionally, we have included mean accuracy (% correct) for the encoding and recognition tasks in Table 2.1, and recognition accuracy across categories can be found in the Supplemental Materials.

Recognition Confidence Scores. First, consistent with our preregistered approach to compare the strength of recognition confidence on old items between conditions, we transformed each response option to a numerical value representing its relative strength (i.e., "definitely old": 4, "probably old": 3, "probably new": 2, and "definitely new": 1). For old trials in the recognition block, we performed a 2 (Visual Noise: visual noise vs. no visual noise) x 3 (Auditory Condition: congruent, control, incongruent) repeated measures analysis of variance (RM ANOVA) on these confidence scores, and post-hoc t-tests with Bonferroni adjusted alpha levels were used for pairwise comparisons. Bayes Factors were also computed for pairwise comparisons to consider the weight of evidence for the tested hypotheses, and interpreted in accordance with Lee & Wagenmakers (2013).

ROC Analysis. The analysis of confidence scores suggests that items belonging to one experimental condition are recognized with higher confidence than those belonging to another group. However, as illustrated in the introduction, an analysis based on hit rates alone is liable to obscure the true nature of latent memory signals. To better characterize the underlying memory signals in each condition, we calculated the hit rates (the proportion of old items correctly identified as old) and false alarm rates (the proportion of new items incorrectly identified as old)

for items in each Auditory Condition at each of our four response options to analyze the underlying ROC (Yonelinas & Parks, 2007). Each subsequent point on an ROC curve relates the hit and false alarm rates as participants increasingly relax their criteria for classifying an item as "old," from "definitely old" to "definitely new." Therefore, the leftmost point of each ROC reflects the hit and false alarm rates for trials on which participants responded "definitely old," the second point from the left reflects the hit and false alarm rates for trials on which participants chose either "definitely old" or "probably old," and so on. The rightmost points have been excluded from the figures because the cumulative hit and false alarm rates converge to one at these points.

For statistical analyses, individual ROCs were constructed for each participant at each level of Auditory Condition, and the points in Figure 2.2b reflect the average observed hit and false alarm rates at each response option across participants. Because we did not observe a significant interaction between Visual Noise and Auditory Condition in our primary analysis, we collapsed across levels of Visual Noise to construct ROCs with a greater number of observations per condition. To compare overall recognition memory strength between Auditory Conditions, we calculated the area under the curve (AUC) of each participants' observed ROCs, which is a theoretically agnostic metric of performance, where a greater area under the curve indicates better recognition memory performance. We performed a RM ANOVA and Bonferroni adjusted post-hoc pairwise t-tests. We note that there were too few response options to fit these ROC data to the dual-process signal detection model, which we address in Experiment 2.

Results

Recognition Confidence Scores. A RM ANOVA showed a significant main effect of Auditory Condition on recognition confidence scores (Figure 2.2a), $F(2, 130) = 8.41$, $p < 0.001$,

η_p^2 = 0.12, such that confidence scores were higher for items encoded in the congruent condition than in the incongruent condition, $t(65) = -3.92$, $p < 0.001$ (see Table 2.1). Bayes Factor indicated very strong evidence for this finding ($BF_{10} = 105.51$). This is consistent with our main hypothesis that stronger memories are formed for visual objects initially paired with congruent compared to incongruent sounds. Post-hoc t-tests did not reveal significant differences between confidence scores for items in congruent and control conditions, or control and incongruent conditions, $t(65) = -1.81$, $p = 0.23$; $t(65) = 2.36$, $p = 0.06$. Nevertheless Bayes Factors did not provide evidence for the null hypotheses for the former comparison ($BF_{01} = 1.59$) and none for the latter ($BF_{01} = 0.56$). There was also a main effect of Visual Noise, $F(1, 65) = 269.05$, $p < 0.001$, $\eta_p^2 = 0.81$, with higher confidence scores for items with no visual noise than with visual noise. There was no significant interaction between Auditory Condition and Visual Noise, $F(2, 130) = 2.72$, $p = 0.07$.

ROC Analysis. A one-way RM ANOVA on AUC for individual ROC curves (Figure 2.2b) revealed a significant effect of Auditory Condition, $F(2, 130) = 5.01$, $p = 0.008$, $\eta_p^2 = 0.07$, post-hoc t-tests showed that memory performance was better for congruent items ($M = 0.82$, $SD = 0.09$) than incongruent items ($M = 0.80$, $SD = 0.10$), $t(65) = 3.02$, $p = 0.01$, with Bayes Factor providing moderate evidence for this finding ($BF_{10} = 8.17$) (Figure 2.2b). AUC was not significantly different between congruent and control items ($M = 0.81$, $SD = 0.10$) or between control and incongruent items, $t(65) = 1.32$, $p = 0.57$; $t(65) = 1.88$, $p = 0.19$, and Bayes Factors

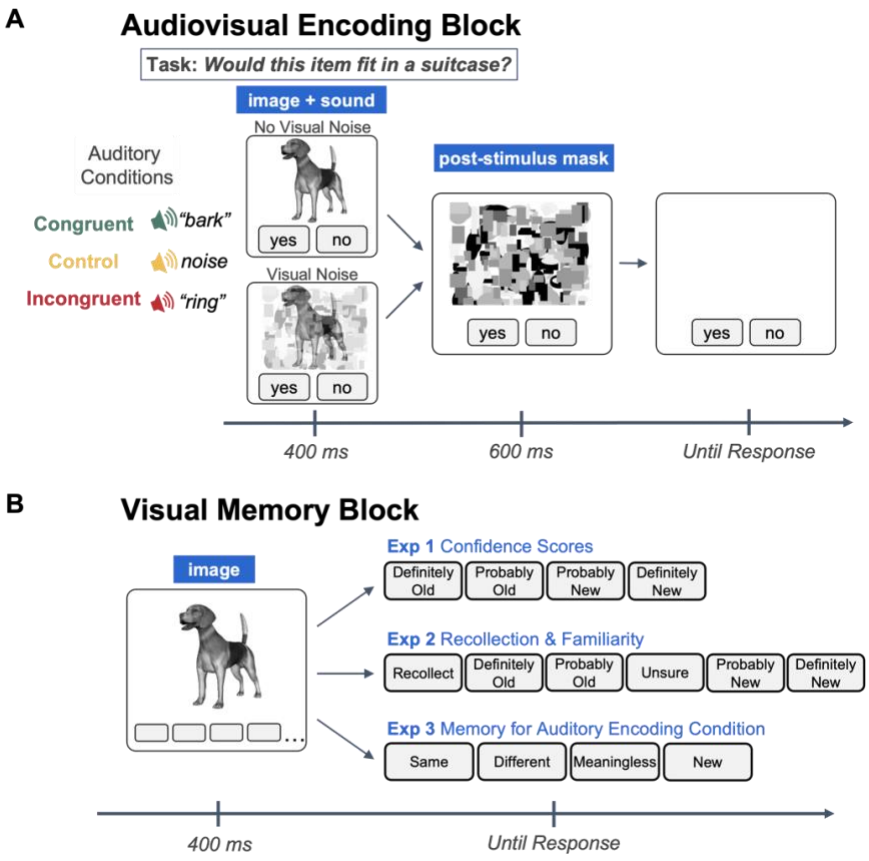


Figure 2.1. A. Illustrates the audiovisual encoding task used for all 3 experiments. Experiments 1 and 3 include the visual noise manipulation during the initial 400ms presentation, while in Experiment 2, all presentations are overlaid by the geometric visual noise during this period. **B.** Surprise visual memory tasks for all 3 experiments.

provide only moderate evidence for the null hypothesis former comparison (BF_{01} = 3.23) and anecdotal for the latter (BF_{01} = 1.41). This pattern of data is consistent with the Confidence Score analysis and shows recognition memory was greater for the congruent than incongruent Auditory Conditions even when response rates are corrected by false alarms at each confidence level. This analysis further illustrates that the difference between conditions decreases as the response criteria relaxes, which indicates that the Auditory Condition primarily affects whether participants recognize items with high confidence.

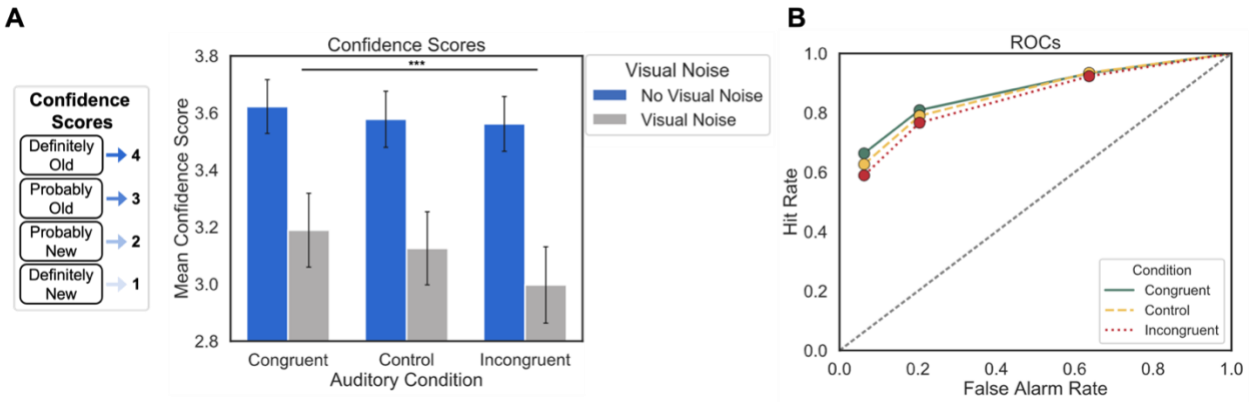


Figure 2.2. A. Mean confidence scores for items in each Visual Noise and Auditory Condition from Experiment 1. The box to the left of the graph illustrates the translation of response options to confidence scores. Error bars denote standard error of the mean. Average confidence scores are higher for items encoded in the congruent than incongruent Auditory Condition, and for items encoded with no visual noise than with visual noise. **B.** Average observed ROCs for each Auditory Condition in Experiment 1, collapsed across Visual Noise Conditions. Each successive point (from left to right) on a given ROC represents the cumulative hit and false alarm rate for items in that condition during the recognition task. Memory performance (AUC) is greater for congruent than incongruent items, and the greatest differences in performance occur at the highest confidence level (leftmost point).

Table 2.1. Mean accuracy for the encoding and recognition tasks and recognition confidence scores for items in each Auditory and Visual Noise condition from Experiment 1. Chance performance is 50% for both encoding and recognition tasks.

	Encoding task accuracy (% correct)		Recognition task accuracy (% correct		Recognition Confidence scores	
Auditory Condition	No Visual Noise	Visual Noise	No Visual Noise	Visual Noise	No Visual Noise	Visual Noise
Congruent	0.84(0.36)	0.79(0.41)	0.86(0.34)	0.72(0.45)	3.62(0.76)	3.18(1.05)
Control	0.84(0.37)	0.79(0.41)	0.86(0.34)	0.70(0.46)	3.58(0.80)	3.13(1.04)
Incongruent	0.81(0.39)	0.75(0.43)	0.85(0.35)	0.66(0.47)	3.56(0.78)	3.00(1.09)

Standard deviations are shown in parentheses.

Experiment 2

The goal of Experiment 2 was to test the hypothesis that congruent sounds, even when they are not relevant to the current task, improve memory by supporting the encoding of details from the episodic event. Results of Experiment 1 suggested that experiencing a visual object in the context of a semantically congruent sound produced better recognition memory than in the context of an incongruent sound, and exploratory analyses suggested that the memory enhancement was specific to the highest level of confidence. Within the dual-process model

framework, such a result could indicate improvement in recollection memory, but not familiarity. This may also explain why significant differences were not detected between the control condition and the congruent or incongruent conditions, because a null effect on familiarity would mitigate an effect driven by recollection-based recognition when the outcome measure includes influences of both. In Experiment 2, we test this hypothesis in a modified recognition task in which the inclusion of additional confidence levels allowed us to obtain formal estimates of recollection and familiarity using the Dual-Process Signal Detection (DPSD) model (Yonelinas, 1994).

Method

Participants. One hundred thirteen students (99 identified as female and 32 identified as male, M_{age} = 20.11) from the University of California, Davis, participated in exchange for partial course credit. Participants were excluded under the same exclusion criteria that were pre-registered criteria for Experiment 1, namely, below chance accuracy on the encoding or recognition task and based on responses to the debriefing survey. We excluded five participants due to low accuracy, and 62 due to debriefing survey responses. The debriefing survey for this experiment included two additional questions regarding comprehension of the recognition task because it was more complex than the task used for Experiment 1. Participants were excluded if they did not fully understand the task or if they did not use the entire range of response options. We used the same target sample size as in Experiment 1, and data was collected until we reached 66 participants post-exclusion.

Materials. All stimuli were identical to those used in Experiment 1.

Procedure. As in Experiment 1, participants completed separate encoding and recognition blocks online via personal computers through the online stimulus presentation software Testable (https://www.testable.org/).

Encoding Block. The encoding block in Experiment 2 was identical to Experiment 1, however, because there was no interaction between Visual Noise Condition and Auditory Condition in Experiment 1, all images in the encoding task for Experiment 2 were overlaid with visual noise instead of half as in Experiment 1. Visual stimuli were counterbalanced across the three, within-subjects Auditory Conditions.

Recognition Block. Immediately following the encoding block, participants completed a visual-only surprise recognition task. This task was almost identical to the recognition task in Experiment 1, except for the response options. On each trial, a visual stimulus was presented for 400 ms, and participants could respond by clicking on buttons corresponding to "recollect," "definitely old," "probably old," "unsure," "probably new," or "definitely new." Participant instructions included a description and example of the difference between a "recollect" response and any "old" response, explaining that "recollect" should only be pressed if the participant was sure that they had seen the item before *and* they could recollect some qualitative information about the encoding event, such as their feelings about the item or what they thought about when they initially saw it.

Debriefing Survey. After the encoding and recognition blocks, participants completed the debriefing survey, which was similar to Experiment 1, and was also used to exclude participants whose testing environment or stimulus presentations were not of adequate quality. To ensure that participants understood the recognition task, the debriefing survey included a question asking whether participants understood when they were supposed to press the "recollect" button, and a

free-response question asking for an example of information they used to judge an item as recollected rather than definitely old (see Supplemental Materials for full list of questions).

Data Analysis. To compare overall memory performance between Auditory Conditions, we calculated AUC from observed ROCs. To directly assess effects of Auditory Conditions on recollection- and familiarity-based recognition, we fit equal variance signal detection models to the observed ROC data in line with the Dual-Process Signal Detection (DPSD) model to compare model parameters associated with these constructs (Yonelinas, 1994). We also analyzed responses to the open-ended debriefing survey prompt asking participants to report an example of information they used to base their "recollect" responses on. Mean accuracy (% correct) across conditions for the encoding and recognition tasks can be found in Table 2.2. It should be noted that "unsure" responses were treated as incorrect for calculating accuracy. Raw data files and analysis code for this experiment are publicly available on the Open Science Framework (https://osf.io/5uz24/).

ROC Analysis. Cumulative hit and false alarm rates were calculated for the observed ROCs just as in Experiment 1, though the response scale was larger for Experiment 2 (in line with previous DPSD studies), so the leftmost point corresponds to the hit and false alarm rates for trials on which participants responded "recollect," the second point from the left reflects the hit and false alarm rates for trials on which participants chose either "recollect" or "definitely old," and so on. For statistical analyses, individual ROCs were constructed for each participant at each level of Auditory Condition, and the points in Figure 2.3a reflect the average observed hit and false alarm rates for these groups across participants. DPSD models were fit to each participants' ROCs as they were in Experiment 1, and the average ROC model for each group is shown in Figure 2.3a. To compare overall differences in memory performance between

conditions, AUC was calculated for each participant's observed ROCs in each Auditory Condition and compared via one-way ANOVA and Bonferroni corrected post-hoc pairwise t-tests. Parameter estimates derived from DPSD model-based ROCs were used to compare two constructs of interest from the dual-process model of recognition memory, namely the *y-intercept*, which estimates recollection, and *d'*, which estimates familiarity. In the DPSD model, the y-intercept estimates the hit rate when the false alarm rate is equal to 0, making it a threshold measure of memory that represents recollection. Model-derived d' measures hit rates relative to false alarm rates across the entirety of the curve, which quantifies the contribution of familiarity. These estimates were also compared via individual one-way ANOVAs and Bonferroni adjusted post-hoc pairwise t-tests. Bayes Factors were also computed for pairwise comparisons and interpreted in the same manner as the previous experiment.

Debriefing Questionnaire Analysis. To perform an exploratory assessment of the details that were recollected about objects on trials for which participants respond with "recollect," we coded the open-ended responses to the debriefing survey for mentions of specific items and/or features recollected from the encoding task (see OSF page for all responses and their categorizations). Responses that referred to objects and their accompanying sound were labeled as "Sound" recollections (e.g., *"I remembered the dog because it was shown along with a 'bark' sound"*). Responses that referred to the objects and other aspects of the encoding experience were labeled as "Not Sound" recollections (e.g., *"I remembered the elephant because elephants are my mom's favorite animal"*). Responses that only listed the name of the visual object were labeled as "Name Only" recollections (e.g., *"bird"*). The responses to each object were summed across the Auditory Condition to which the object was encoded for each participant. In cases where a participant mentioned multiple items, all items were included in the analysis, so the total

number of responses exceeds the number of items included in the analysis. For the analysis, we compared the count of items mentioned from each Auditory Condition (congruent, incongruent, control), and the detail given as part of the response ("Sound," "Not Sound," and "Name Only"). A small proportion of the responses included a detail that was recollected without mentioning a specific item (e.g., *"I pressed "recollect" if I remembered the sound that played with an item in the first task"*). These responses are categorized as "Nonspecific" because they do not contain explicit object labels and are discussed separately from the responses that did mention specific items, and are not included in figure 2.3d.

Results

ROC analysis.

AUC. A one-way RM ANOVA revealed a significant effect of Auditory Condition on AUC, $F(2, 130) = 9.84$, $p < 0.001$, $\eta_p^2 = 0.13$, and post-hoc t-tests showed that memory performance was better for congruent items than control or incongruent items, $t(65) = 3.94$, $p < 0.001$; $t(65) = 3.86$, $p < 0.001$, and Bayes Factors suggest very strong evidence for both findings ($BF_{10} = 112.20$; $BF_{10} = 86.51$). (Table 2). AUC was not significantly different between control and incongruent items, $t(65) = -0.41$, $p = 1.00$, and Bayes Factor provide moderate evidence for the null hypothesis ($BF_{01} = 6.67$). These results show that recognition was better for items in the congruent condition than the control or incongruent conditions.

Recollection and Familiarity. A one-way RM ANOVA showed a significant effect of Auditory Condition on recollection (*y-intercept*), $F(2, 130) = 11.10$, $p < 0.001$, $\eta_p^2 = 0.15$ (Figure 2.3b), with higher y-intercepts for items in the congruent condition than in the control condition or incongruent condition, but no significant difference between items in the control and incongruent conditions, $t(65) = 4.03$, $p < 0.001$; $t(65) = 4.65$, $p < 0.001$; $t(65) = -0.20$, $p = 1.00$,

and Bayes Factors provided very strong evidence for the findings in the first two comparisons (BF_{10} = 148.66; BF_{10} = 1084.05), and moderate evidence for the null hypothesis in the latter (BF_{01} = 7.14). A one-way RM ANOVA did not show a significant effect of Auditory Condition on familiarity (*d'*), $F(2, 130) = 1.02$, $p = 0.37$, $\eta_p^2 = 0.02$ (Figure 2.3c). These results confirm our hypothesis and converge with the exploratory analysis in Experiment 1, showing that improvements in memory for the congruent Auditory Condition were due to better recollection-based recognition memory. Interestingly, we found no effect of Auditory Condition on familiarity, suggesting that the effect of an auditory event was specific to encoding mechanisms that improve recollection.

Debriefing Questionnaire. Out of a total of 76 responses, ten (13.2%) were "Nonspecific," and of these, 3 mentioned that they chose "recollect" if they remembered the sound that an item was paired with but neither the object nor the sound was explicitly named; the other 7 mentioned non-sound details, but also did not provide explicit labels for the objects. The 66 responses that did specifically name items included 34 named congruent items, 19 control items, and 13 incongruent items. 50% of the named congruent items included "Sound" details, while only 10.5% of control items mentioned "Sound" details, and none of the incongruent items included "Sound" details (Figure 2.3d). There were a similar number of responses in each of the three conditions mentioning "Not Sound" details (Figure 2.3d).

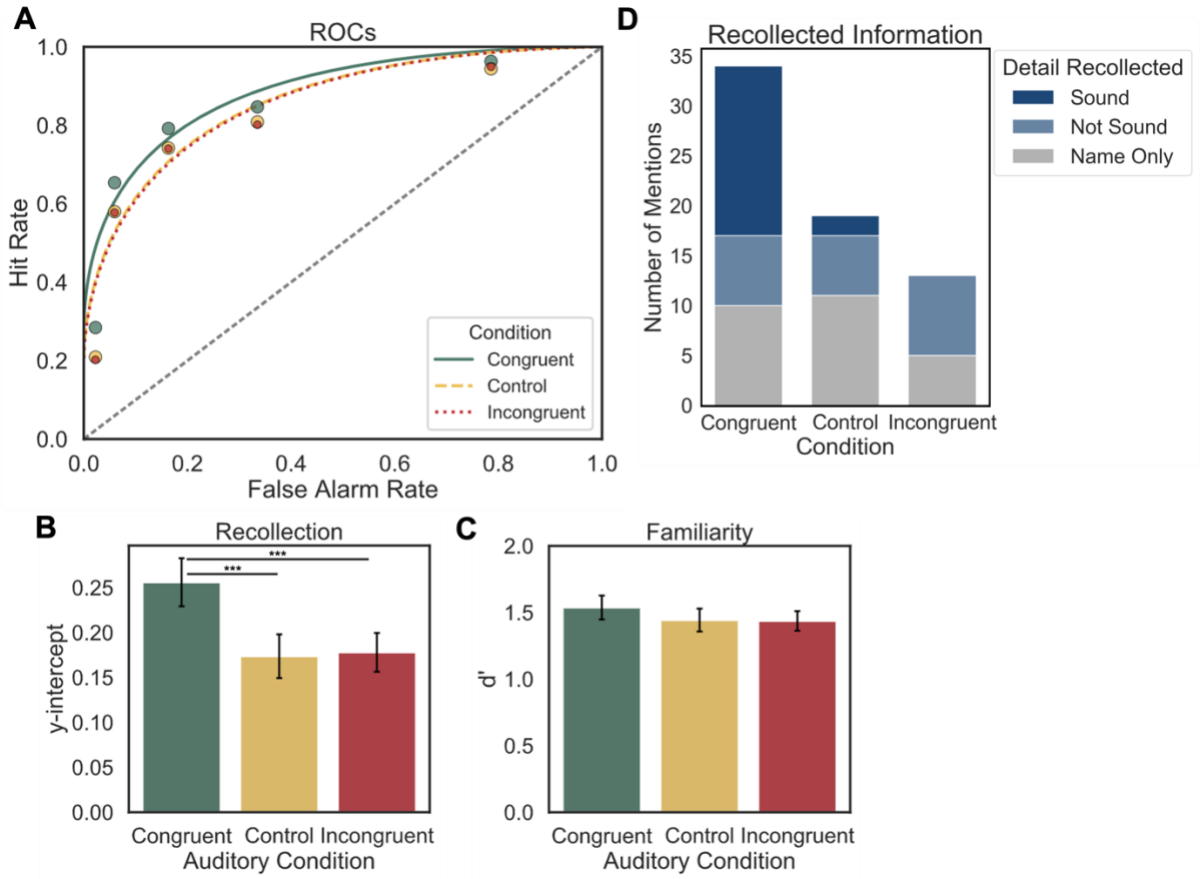


Figure 2.3. A. The average observed ROCs (points) for each Auditory Condition from Experiment 2 and corresponding DPSD equal-variance signal detection model functions. Overall memory performance (AUC) is greater for congruent than incongruent or control items. **B.** Average *y-intercept* for DPSD ROC curves between Auditory Conditions. Recollection is greater for congruent than incongruent or control items. **C.** DPSD model-derived *d'* for each Auditory Condition. No significant differences between conditions. **D.** Mentions of each type of recollected detail from the debriefing survey for items from each condition. Responses that did not mention a specific item ("Nonspecific" responses) are not included in the figure. All error bars denote standard error.

Table 2.2. Mean accuracy for the encoding and recognition tasks, and mean DPSD model parameters for overall recognition memory (*AUC*), recollection (y-intercept), and familiarity (*d'*) for items in each condition for Experiment 2. Chance performance is 50% accuracy for both encoding and recognition tasks.

Auditory Condition	Task Accuracy (% correct)		DPSD Model Parameters		
	Encoding Task	Recognition Task	*AUC*	*y-intercept*	*d'*
Congruent	0.81(0.39)	0.79(0.41)	0.85(0.09)	0.26(0.22)	1.54(0.73)
Control	0.80(0.40)	0.74(0.44)	0.82(0.10)	0.17(0.20)	1.44(0.70)
Incongruent	0.81(0.39)	0.74(0.44)	0.82(0.10)	0.18(0.18)	1.44(0.60)

Standard deviations are shown in parentheses.

Experiment 3

Experiment 2 found that a semantically congruent multisensory event led to better recollection-based recognition memory, indicating that they produced a more detailed memory of the encoded event. Based on this finding, we would expect not only better recognition of the visual object, but also better memory for the association between the visual object and the sound. We tested this hypothesis in Experiment 3 by altering the memory test to ask participants in which Auditory Condition they experienced each visual object. Although the recall task was expected to be more difficult, fewer items were included in this experiment, and therefore two levels of Visual Noise were included to prevent possible ceiling or floor effects.

Method

Participants. Seventy-six students (65 identified as female, 10 identified as male, and one identified as other M_{age} = 19.17) from the University of California, Davis, participated in exchange for partial course credit. Participants were excluded under our pre-registered exclusion criteria, namely, below chance accuracy on the encoding or recognition task and based on responses to the debriefing survey. Ten participants were excluded due to low accuracy, and zero due to debriefing survey responses. We used the same target sample size as in Experiments 1 and 2, and data was collected until we reached 66 participants post-exclusion.

Materials. All stimuli were identical to those used in Experiments 1 and 2. However, in the recognition task, instead of 90 new items, there were only 30 in order to keep the number of items equal across each of the four response options, for a total of 120 items in the recognition task (see Supplemental Materials).

Procedure. As in Experiments 1 and 2, participants completed separate encoding and recognition blocks online via personal computers through the online stimulus presentation software Testable (https://www.testable.org/).

Encoding Block. The encoding task in Experiment 3 was identical to the encoding task in Experiment 1. We included the visual noise manipulation from Experiment 1 as a precaution to ensure that ceiling or floor effects would be avoided (see Figure 2.1a).

Recognition Block. Immediately following the encoding block, participants completed a visual-only surprise recognition task. This task was similar to the tasks in Experiments 1 and 2, with a few exceptions. 90 "old" items were mixed with 30 "new" items, and on each trial, participants were asked to indicate whether the object was originally presented with a sound that was the same as the object (congruent), different from the object (incongruent), a meaningless, white-noise sound (control), or if the object was new (see Figure 2.1b).

Debriefing Survey. After the encoding and recognition blocks, participants completed the debriefing survey, which was the same as the survey used for Experiment 1 and was also used to exclude participants whose testing environment or stimulus presentations were not of adequate quality (see Supplemental Materials for full list of questions).

Data Analysis. To assess memory for the auditory encoding condition in Experiment 3, we calculated the sensitivity index *d'* for hits and false alarms for old items in each Auditory Condition. In this experiment, a hit occurred when an old item was attributed to the correct encoding condition (congruent, control, or incongruent sound), and a false alarm occurred when an old item was attributed to the incorrect encoding condition. Our preregistered analysis plan included a RM ANOVA to compare raw memory accuracy (percent correct) between Auditory and Visual Noise Conditions, though we deviated from this plan because the false alarms in each condition were unevenly distributed across response options. Specifically, when participants saw an old item that had initially been presented in the control or incongruent conditions, they most often incorrectly attributed these items as belonging to the congruent condition during encoding.

As such, this potentially inflated the raw accuracy of the congruent condition, so we used the measure of *d'* to avoid this potential confound. It should be noted that new items had false alarms that were evenly distributed across the congruent, control, and incongruent responses, suggesting that the response bias was unique to old items. We performed a 2 (Visual Noise: visual noise vs. no visual noise) x 3 (Auditory Condition: congruent, control, incongruent) RM ANOVA on the *d'* performance index, and Bonferroni adjusted post-hoc pairwise t-tests. Bayes Factors were computed for pairwise comparisons. Additionally, mean accuracy (% correct) for the encoding and recognition tasks can be found in Table 2.3.

Results

A RM ANOVA on memory for the encoding condition (*d'*) showed no significant interaction between Visual Noise and Auditory Condition, $F(2, 130) = 0.98$, $p = 0.38$, and no significant effect of Visual Noise, $F(1, 65) = 0.41$, $p = 0.53$, $\eta_p^2 = 0.006$. However, there was a significant effect of Auditory Condition, such that memory for the auditory encoding condition was better for items in the congruent encoding condition than the control or incongruent conditions, but no difference between control and incongruent conditions, $F(2, 130) = 29.95$, $p < 0.001$, $\eta_p^2 = 0.32$; $t(65) = 6.18$, $p < 0.001$; $t(65) = 6.78$, $p < 0.001$; $t(65) = -0.85$, $p = 1.00$, with Bayes Factors provided very strong evidence for the finding from the first two comparisons (BF_{10} = 2.73x10^5, BF_{10} = 2.73x10^6), and moderate evidence for the null hypothesis for the last comparison (BF_{01} = 5.26) (Table 3, Figure 2.4). These results suggest that in addition to better memory for the visual stimulus, the presence of a congruent sound facilitates the retrieval of the task-irrelevant auditory stimulus, even though participants were not aware that their memory would be tested for either.

Table 2.3. Mean accuracy for the encoding and recognition tasks and mean *d'* for each Auditory Condition and Visual Noise Condition from Experiment 3. Chance performance is 50% accuracy for the encoding and 25% accuracy for the recognition task.

	Encoding Task (% Correct)		Recognition Task (% Correct)		Recognition Task *d'*	
Auditory Condition	No Visual Noise	Visual Noise	No Visual Noise	Visual Noise	No Visual Noise	Visual Noise
Congruent	0.89(0.31)	0.82(0.38)	0.65(0.48)	0.55(0.50)	0.97(0.66)	0.88(0.58)
Control	0.88(0.32)	0.79(0.41)	0.22(0.42)	0.19(0.40)	0.52(0.48)	0.56(0.42)
Incongruent	0.85(0.36)	0.78(0.41)	0.28(0.45)	0.24(0.43)	0.51(0.46)	0.47(0.40)

Standard deviations are shown in parentheses.

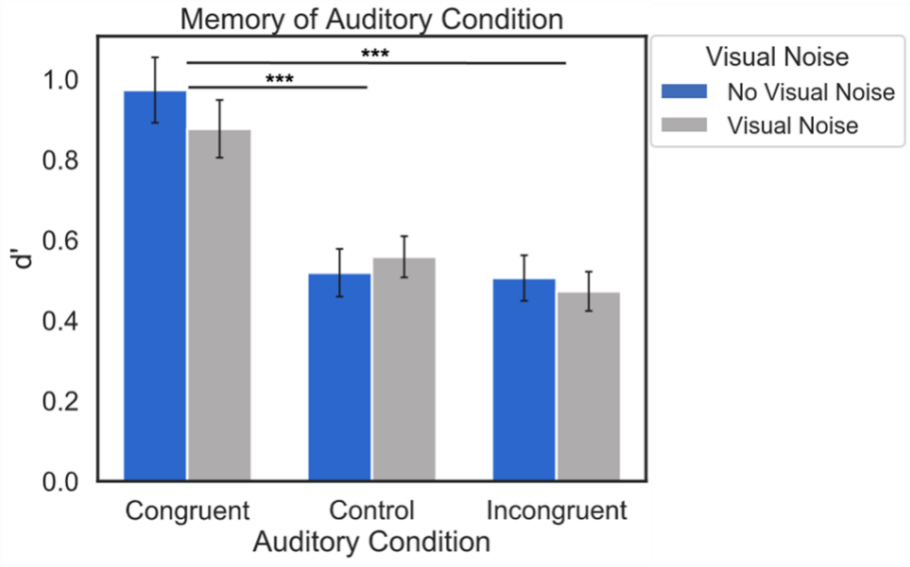


Figure 2.4. Average *d'* for each Visual Noise and Auditory Condition. Error bars denote standard error of the mean. Memory was better for the sound played during encoding for items in the congruent condition than the incongruent or control conditions.

Discussion

The goal of the present research was to investigate whether congruent multisensory presentation facilitates visual recognition memory by supporting recollection or familiarity-based recognition. Our results replicated previous findings (Lehmann & Murray, 2005; Thelen et al., 2015), even with tests of incidental memory and when hit rates were compared across multiple false alarm rates. More importantly, consistent with our hypothesis, our results provide the first evidence that memory improvement for semantically congruent audiovisual pairs specifically

promotes recollection. We also showed that learning object-congruent sounds not only improved memory for the task-relevant visual object, but also for the sounds themselves despite being task-irrelevant and the memory tests being completely unexpected. Together, our experiments demonstrate that the presence of an object-congruent sound at encoding increases the likelihood that an episodic memory for an object will be formed and later recollected.

Our findings also suggest that this memory benefit is due to the integration of semantically congruent information into the encoded object representation, and that improvement to perceptual fluency during encoding cannot alone explain our findings. In Experiment 1, visual noise impaired encoding and recognition performance overall, but the impact was equivalent for congruent, incongruent, and control audiovisual object pairs. If the benefit for audiovisual pairs in recognition memory stemmed from increased perceptual fluency, the effect of visual noise should have been smaller for congruent pairs, but that was not the case. However, we cannot rule out the possibility that the specific conditions used in our experiments may have reduced the effect of perceptual fluency. For example, Chen & Spence (2010) used much briefer image presentations (27 ms) than we did (400 ms), and found benefits of multisensory processing on perceptual fluency for object identification (see also Driver & Noesselt, 2008). Experiment 3 similarly showed overall poorer encoding and recognition accuracy with visual noise across conditions, and there was no main effect of Visual Noise when accounting for response biases using *d'*. Overall, the lack of interaction between Auditory Condition and Visual Noise in these studies suggests that multisensory processing did not facilitate recognition memory merely by increasing perceptual fluency.

Experiment 2 suggests that the multisensory memory effect is driven by a mechanism that facilitates the storage, and later recollection, of details from the encoded event, particularly

for the sound itself. This extends previous research on semantic congruency of word-pairs on recollection—rather than familiarity—to more naturalistic stimuli (Bein et al., 2015). In both the Bein et al. (2015) study and ours, the recollection benefit may be specific to retrieval of the accessory information presented at encoding (the adjective in their study and the sound in the present study). Analysis of the debriefing survey in Experiment 2 supports this possibility, and future research will be needed to investigate whether a congruent sound at encoding leads to better recollection because it helped reinstate visual object details or the context, or because it reduced processing demands associated with encoding consistent information along for more visual details to be encoded in the first place. Evidence that redundant multisensory signals provide neural and behavioral benefits over redundant unisensory signals in cats (Alvarado et al., 2007; Gingras et al., 2009) and humans (Laurienti et al., 2004) suggests that crossmodal accessory information could provide support over-and-above accessory unisensory information, but such comparisons have yet to be made in memory studies. Regardless of whether there is more than one mechanism underlying the multisensory memory benefit, the present study presents an ecologically valid situation in which congruent semantic information facilitates later recollection-based memory.

Experiment 3 showed that the relation between the objects and sounds was more likely to be remembered when the sound was congruent, despite being task-irrelevant during encoding. This is consistent with research showing that attention spreads from a task-relevant to a task-irrelevant stimulus in another modality that corresponds semantically (Fiebelkorn et al., 2010; Molholm et al., 2007; Zimmer et al., 2010). For example, Molholm and colleagues (2007) showed serial audiovisual presentations of objects that were congruent or incongruent, and had participants perform an N-back task on stimuli in either the visual or auditory modality.

Processing of the stimulus in the ignored modality was enhanced when it was semantically congruent to the stimulus in the attended modality, as indexed by the SN ERP component. In the current studies, attention likely also spread from visual items to semantically congruent sounds, and this attentional enhancement of multiple pieces of object-related information may be responsible for the recollection-specific benefit (Greene et al., 2021; Craik et al., 1996; Troyer & Craik, 2000)[1].

This interpretation fits well within a predictive coding framework, which posits that the brain maintains an internal model of the environment that generates predictions about the environment and sensory inputs are either confirmatory or produce an error signal (Friston, 2010; Friston & Kiebel, 2009). Talsma (2015) proposed that congruent crossmodal stimuli produce a signal with low prediction error, resulting in a stronger memory trace and less effortful encoding than an incongruent crossmodal stimulus that would produce an error signal and require a model update. Such a mechanism could explain our results because attention to two congruent constituents of a multisensory stimulus would be expected and reinforce the same internal representation, leading the object to be more readily bound into an episodic memory than if attention is divided when the audiovisual event is incongruent[2].

In summary, the present studies expand upon research on the memory benefits of congruent multisensory events by showing that a visual object encoded with a congruent sound is more likely to be recognized later based on detailed recollection. While our evidence indicates

[1] It is possible that this enhanced attention to object-specific sensory information could specifically facilitate memory for that visual object and sound itself, thereby incurring a cost to memory for other aspects of the encoded event, though future studies will be needed to understand whether the present effect comes at any sort of impairment.

[2]Across our experiments, we did not find memory impairments for items presented with incongruent sounds relative to control sounds as previous studies have, which may have been due to the relatively high proportion of incongruent trials included in our experiments (see Thelen et al., 2015).

that memory for the sound supports recollection-based recognition of the visual object, future studies will be necessary to determine whether memory for other details, such as specific visual features or its context are also improved. Nevertheless, the present study illustrates how multisensory events can produce a qualitative shift in the encoding of episodic events.

Chapter 2 Supplemental Materials

Supplemental Table 2.1. Visual object stimuli for Experiments 1-3.

Experiment	Old/New	Size	Object	Category
1, 2, 3	Old	Small	hammer	tool
			saw	tool
			bird	animal
			hen	animal
			frog	animal
			rat	animal
			cat	animal
			bat	animal
			smartphone	common
			keyboard	common
			bell	instrument
			snake	animal
			stapler	household
			scissors	household
			camera	common
			drill	tool
			basketball	recreation
			blender	household
			book	household
			bowling pin	recreation
			chick	animal
			clock	household
			cup	household
			straw	household
			flute	instrument
			coins	common
			laptop	common
			hairdryer	household
			harmonica	instrument
			kettle	household
			key	household
			lighter	tool
			light switch	household
			maracas	instrument
			matches	tool
			pencil	household
			Ping-Pong paddle	recreation
			tape	household
			soda can	household
			spray bottle	household
			teapot	household
			tennis racket	recreation
			toaster	household
			wine bottle	household
			xylophone	instrument
1, 2, 3	Old	Large	elephant	animal
			dog	animal

			horse	animal
			pig	animal
			cow	animal
			bear	animal
			tiger	animal
			motorcycle	vehicle
			guitar	instrument
			jet	vehicle
			car	vehicle
			piano	instrument
			drum	instrument
			cymbal ride	instrument
			goat	animal
			deer	animal
			leopard	animal
			wolf	animal
			arcade game	recreation
			axe	tool
			baseball bat	recreation
			billiards	recreation
			bicycle	vehicle
			train	vehicle
			door	common
			washer	household
			crocodile	animal
			goose	animal
			penguin	animal
			boat	vehicle
			microwave	household
			printer	common
			sink	household
			skateboard	vehicle
			sled	recreation
			sword	common
			toilet	household
			toilet brush	household
			Bow/arrow	recreation
			chair	household
			anvil	tool
			filing cabinet	household
			fireplace	household
			golfclub	recreation
			helicopter	vehicle
1,2,3	New	Mixed	wrench	tool
			whale	animal
			watering can	household
			walkie	common
			scooter	vehicle
			shark	animal
			salamander	animal
			record player	instrument

			rhino	animal
			rabbit	animal
			fridge	household
			fish	animal
			crab	animal
			frypan	household
			camel	animal
			butterfly	animal
			bus	vehicle
			radio	common
			air hockey	recreation
			briefcase	common
			calculator	common
			candle	household
			football	recreation
			hoe	tool
			lamp	household
			microphone	instrument
			notepad	common
			plant	household
			pliers	tool
			scorpion	animal
1, 2	New	Mixed	pen	common
			screwdriver	tool
			snail	animal
			soccer ball	recreation
			spider	animal
			spoon	common
			table	household
			turtle	animal
			umbrella	common
			water bottle	common
			whisk	household
			eraser	common
			octopus	animal
			screw	tool
			toucan	animal
			zebra	animal
			bed	household
			cake stand	household
			closet	household
			couch	household
			espresso maker	household
			tree	common
			stove	household
			barrel	common
			bucket	common
			crowbar	tool
			rope	tool
			scale	tool
			tv	common

			bedside table	household
			mirror	common
			mocha pot	household
			bathtub	household
			spatula	household
			basket	common
			box	common
			duffel	common
			giftbox	common
			pallet	tool
			shelf	common
			suitcase	common
			trash can	common
			treasure chest	common
			extinguisher	tool
			tire	common
			chess	recreation
			cone	common
			streetlight	common
			hydrant	tool
			ladder	tool
			mailbox	common
			street sign	common
			beach chair	recreation
			fork	household
			life ring	tool
			outlet	common
			shovel	tool
			surfboard	recreation
			clipboard	common
			globe	common

Supplemental Table 2.2. Incongruent sounds used across all 3 experiments, and the objects they were paired with in each counterbalanced version of the experiment.

Sound	**Incongruent Visual Pairing 1**	**Incongruent Visual Pairing 2**	**Incongruent Visual Pairing 3**
ambulance	drill	key	hammer
bagpipe	basketball	lighter	saw
bee	blender	light switch	bird
bongo	book	maracas	hen
broom	bowling pin	matches	frog
chainsaw	chick	pencil	rat
chimp	clock	ping pong paddle	cat
comb	cup	microwave	elephant
cricket	deer	printer	dog
donkey	leopard	sink	horse
fencing	wolf	skateboard	pig
fire alarm	arcade game	sled	cow
foghorn	axe	sword	bear
glass	baseball bat	toilet	tiger
gong	billiards	toilet brush	motorcycle

grate	cup straw	tape	bat
guinea pig	flute	soda can	smartphone
gun	coins	spray bottle	keyboard
harp	laptop	teapot	bell
ice cube	hairdryer	tennis racket	snake
morse	harmonica	toaster	stapler
music box	kettle	wine bottle	scissors
razor	bicycle	xylophone	camera
roulette	train	bow arrow	guitar
snapper	door	chair	jet
sealion	washer	anvil	car
trampoline	crocodile	filing cabinet	piano
trumpet	goose	fireplace	drum
whip	penguin	golfclub	cymbal ride
whistle	boat	helicopter	goat

Supplemental Table 2.3. The number of old and new exemplars per category and the recognition accuracy (% correct) for Experiments 1 and 2. Note that "unsure" responses in Experiment 2 were considered incorrect.

Category	Old Items	New Items	Difference	Recognition Accuracy Exp 1	Recognition Accuracy Exp 2
common misc.	8	31	-23	0.81(39)	0.69(0.46)
tool	7	14	-7	0.77(42)	0.63(0.45)
recreation	10	6	4	0.87(34)	0.82(0.38)
animal	22	16	6	0.81(0.40)	0.77(0.42)
vehicle	8	2	6	0.76(43)	0.72(0.45)
household	26	19	7	0.76(42)	0.67(0.47)
instrument	9	2	7	0.81(39)	0.82(0.38)

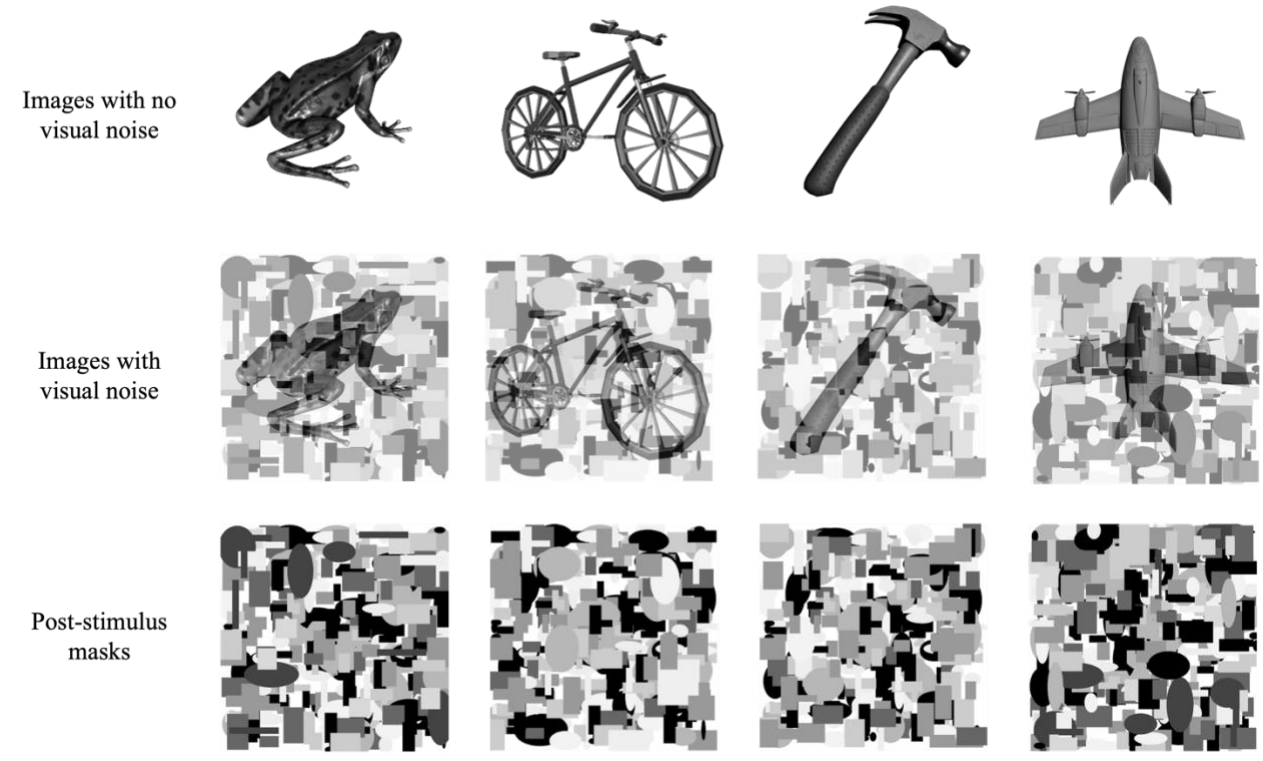
*Standard deviations are shown in parentheses

Supplemental Table 2.4. Debriefing questions for experiments 1-3

Experiment	Question	Response Options
1, 2, 3	Was the volume on your computer enabled throughout the entire first task?	Yes; No
	Did you adjust your volume at any time during the experiment?	Yes; No
	How would you describe the volume of the sounds during the first task?	Quiet; Loud; Just right; I did not hear sounds
	Did you use external speakers, in-ear headphones, over-ear headphones?	External speakers; In-ear headphones; Over-ear headphones; Other; I did not hear sounds
	How would you describe the environment in which you completed the study?	Quiet; Mostly quiet; Somewhat noisy; Very noisy
	Did you experience distractions during the study?	Yes, major distractions; Yes, minor distractions; No
	How much effort did you put into the experiment?	Not any; Not very much; Some effort; A lot of effort

	How difficult did you find task 1?	Not difficult; A little difficult; Very difficult
	How difficult did you find task 2?	Not difficult; A little difficult; Very difficult
	In the first task, did you ever experience a lag or gap between when the picture was shown and when the sound started?	Yes, a couple of times; Yes, often or always; No; I did not hear sounds
2	In the second task, did you understand when you were supposed to press the "Recollect" button?	Yes, definitely; Yes, I think so; Not sure; Not at all
	Please give an example of something you recalled about an object on a trial where you pressed the "Recollect" button.	Free-response

Supplemental Table 2.5. Example visual stimuli. In the encoding tasks, images with and without visual noise were used for Experiments 1 and 3, and only images with visual noise were used for Experiment 2. Post-stimulus masks were presented immediately after images in the encoding task across experiments. In the memory test of all three experiments, the images were never overlaid with visual noise.



References

Alvarado, J. C., Vaughan, J. W., Stanford, T. R., & Stein, B. E. (2007). Multisensory Versus Unisensory Integration: Contrasting Modes in the Superior Colliculus. *Journal of Neurophysiology*, *97*(5), 3193–3205. https://doi.org/10.1152/jn.00018.2007

Atienza, M., Crespo-Garcia, M., & Cantero, J. L. (2011). Semantic Congruence Enhances Memory of Episodic Associations: Role of Theta Oscillations. *Journal of Cognitive Neuroscience*, *23*(1), 75–90. https://doi.org/10.1162/jocn.2009.21358

Bein, O., Livneh, N., Reggev, N., Gilead, M., Goshen-Gottstein, Y., & Maril, A. (2015). Delineating the Effect of Semantic Congruency on Episodic Memory: The Role of Integration and Relatedness. *PLOS ONE*, *10*(2), e0115624. https://doi.org/10.1371/journal.pone.0115624

Brady, T. F., Robinson, M. M., Williams, J. R., & Wixted, J. (2021). *Measuring memory is harder than you think: A crisis of measurement in memory research* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/qd75k

Chen, Y.-C., & Spence, C. (2010). When hearing the bark helps to identify the dog: Semantically-congruent sounds modulate the identification of masked pictures. *Cognition*, *114*(3), 389–404. https://doi.org/10.1016/j.cognition.2009.10.012

Diana, R. A., Yonelinas, A. P., & Ranganath, C. (2007). Imaging recollection and familiarity in the medial temporal lobe: A three-component model. *Trends in Cognitive Sciences*, *11*(9), 379–386. https://doi.org/10.1016/j.tics.2007.08.001

Driver, J., & Noesselt, T. (2008). Multisensory Interplay Reveals Crossmodal Influences on 'Sensory-Specific' Brain Regions, Neural Responses, and Judgments. *Neuron*, *57*(1), 11–23. https://doi.org/10.1016/j.neuron.2007.12.013

Edmiston, P., & Lupyan, G. (2015). What makes words special? Words as unmotivated cues. *Cognition, 143,* 93–100. https://doi.org/10.1016/j.cognition.2015.06.008

Fiebelkorn, I. C., Foxe, J. J., & Molholm, S. (2010). Dual Mechanisms for the Cross-Sensory Spread of Attention: How Much Do Learned Associations Matter? *Cerebral Cortex*, *20*(1), 109–120. https://doi.org/10.1093/cercor/bhp083

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1211–1221. https://doi.org/10.1098/rstb.2008.0300

Gingras, G., Rowland, B. A., & Stein, B. E. (2009). The Differing Impact of Multisensory and Unisensory Integration on Behavior. *Journal of Neuroscience*, *29*(15), 4897–4902. https://doi.org/10.1523/JNEUROSCI.4120-08.2009

Greene, N. R., Martin, B. A., & Naveh-Benjamin, M. (2021). The effects of divided attention at encoding and at retrieval on multidimensional source memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0001051

Hashtroudi, S. (1983). Type of semantic elaboration and recall. *Memory & Cognition*, *11*(5), 476–484. https://doi.org/10.3758/BF03196984

Heikkilä, J., Alho, K., Hyvönen, H., & Tiippana, K. (2015). Audiovisual Semantic Congruency During Encoding Enhances Memory Performance. *Experimental Psychology*, *62*(2), 123–130. https://doi.org/10.1027/1618-3169/a000279

Laurienti, PaulJ., Kraft, RobertA., Maldjian, JosephA., Burdette, JonathanH., & Wallace, MarkT. (2004). Semantic congruence is a critical factor in multisensory behavioral performance. *Experimental Brain Research*, *158*(4). https://doi.org/10.1007/s00221-004-1913-2

Lee, M. D., & Wagenmakers, E.-J. (2013). Bayesian Cognitive Modeling: A Practical Course. Cambridge University Press. https://doi.org/10.1017/CBO9781139087759

Lehmann, S., & Murray, M. M. (2005). The role of multisensory memories in unisensory object discrimination. *Cognitive Brain Research*, *24*(2), 326–334. https://doi.org/10.1016/j.cogbrainres.2005.02.005

Matusz, P. J., Wallace, M. T., & Murray, M. M. (2017). A multisensory perspective on object memory. *Neuropsychologia*, *105*, 243–252. https://doi.org/10.1016/j.neuropsychologia.2017.04.008

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: Multisensory transfer of object-based attention. *European Journal of Neuroscience*, *26*(2), 499–509. https://doi.org/10.1111/j.1460-9568.2007.05668.x

Moran, Z. D., Bachman, P., Pham, P., Hah Cho, S., Cannon, T. D., & Shams, L. (2013). Multisensory Encoding Improves Auditory Recognition. *Multisensory Research*, *26*(6), 581–592. https://doi.org/10.1163/22134808-00002436

Murray, M. M., Foxe, J. J., & Wylie, G. R. (2005). The brain uses single-trial multisensory memories to discriminate without awareness. *NeuroImage*, *27*(2), 473–478. https://doi.org/10.1016/j.neuroimage.2005.04.016

Murray, M. M., Michel, C. M., Grave de Peralta, R., Ortigue, S., Brunet, D., Gonzalez Andino, S., & Schnider, A. (2004). Rapid discrimination of visual and multisensory memories

revealed by electrical neuroimaging. *NeuroImage*, *21*(1), 125–135. https://doi.org/10.1016/j.neuroimage.2003.09.035

Shams, L., & Seitz, A. R. (2008). Benefits of multisensory learning. *Trends in Cognitive Sciences*, *12*(11), 411–417. https://doi.org/10.1016/j.tics.2008.07.006

Stein, B. E., Stanford, T. R., & Rowland, B. A. (2020). Multisensory Integration and the Society for Neuroscience: Then and Now. *The Journal of Neuroscience*, *40*(1), 3–11. https://doi.org/10.1523/JNEUROSCI.0737-19.2019

Talsma, D. (2015). Predictive coding and multisensory integration: An attentional account of the multisensory mind. *Frontiers in Integrative Neuroscience*, *09*. https://doi.org/10.3389/fnint.2015.00019

Thelen, A., Talsma, D., & Murray, M. M. (2015). Single-trial multisensory memories affect later auditory and visual object discrimination. *Cognition*, *138*, 148–160. https://doi.org/10.1016/j.cognition.2015.02.003

Wais, P. E., Wixted, J. T., Hopkins, R. O., & Squire, L. R. (2006). The Hippocampus Supports both the Recollection and the Familiarity Components of Recognition Memory. *Neuron*, *49*(3), 459–466. https://doi.org/10.1016/j.neuron.2005.12.020

Yonelinas, A. P. (2002). The Nature of Recollection and Familiarity: A Review of 30 Years of Research. *Journal of Memory and Language*, *46*(3), 441–517. https://doi.org/10.1006/jmla.2002.2864

Yonelinas, A. P., Aly, M., Wang, W.-C., & Koen, J. D. (2010). Recollection and familiarity: Examining controversial assumptions and new directions. *Hippocampus*, *20*(11), 1178–1194. https://doi.org/10.1002/hipo.20864

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin*, *133*(5), 800–832. https://doi.org/10.1037/0033-2909.133.5.800

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

Chapter 3: Multisensory processing impacts memory for objects and their sources

The following chapter consists of a manuscript that has been published in the journal *Memory & Cognition*

Abstract

Multisensory object processing improves recognition memory for individual objects, but its impact on memory for neighboring visual objects and scene context remains largely unknown. It is therefore unclear how multisensory processing impacts episodic memory for information outside of the object itself. We conducted three experiments to test the prediction that the presence of audiovisual objects at encoding would improve memory for nearby visual objects, and improve memory for the environmental context in which they occurred. In Experiments 1a and 1b, participants viewed audiovisual-visual object pairs or visual-visual object pairs with a control sound during encoding and were subsequently tested on their memory for each object individually. In Experiment 2, objects were paired with semantically congruent or meaningless control sounds and appeared within four different scene environments. Memory for the environment was tested. Results from Experiments 1a and 1b showed that encoding a congruent audiovisual object did not significantly benefit memory for neighboring visual objects, but Experiment 2 showed that encoding a congruent audiovisual object did improve memory for the environments in which those objects were encoded. These findings suggest that multisensory processing can influence memory beyond the objects themselves and that it has a unique role in episodic memory formation. This is particularly important for understanding how memories and associations are formed in real-world situations, in which objects and their surroundings are often multimodal.

Introduction

The ability to remember events is fundamental to the human experience. In order to successfully store information into episodic memory, the brain must form novel associations between individual objects (e.g., a sea lion and seagulls) and their context (e.g., San Francisco's Pier 39; (Mitchell & Johnson, 2009; Dickerson & Eichenbaum, 2010; Yonelinas et al., 2019). While memories can be formed for individual objects, episodic memories are unique in that they consist of relational representations including spatial, temporal, and other details of the event in which the objects were encountered to provide a cohesive story about an experience. Laboratory studies of episodic memories have mostly focused on visual stimuli, e.g., seeing a sea lion at Pier 39, but real-world experiences involve information presented through multiple senses at the same time: the sea lion *barks* and the seagulls *squawk*. In contrast, there have been many studies of multisensory object processing, but most of these have focused on the processing of individual objects that appear alone. These studies show that multisensory objects are identified faster, are more likely to be attended and remembered, and facilitate learning (Shams & Seitz, 2008; Macaluso et al., 2016; Matusz et al., 2017; Stein et al., 2020). However, little is known about how multisensory objects impact the formation of episodic memories. The current experiments bring together two literatures to test the hypotheses that the presence of a multisensory object at encoding will improve memory for nearby visual objects and the environmental context.

A number of studies have demonstrated that general recognition memory is better for visual objects encoded with an object-congruent sound (e.g., dog + *bark*) compared to an incongruent sound (e.g., dog + *ding*), a meaningless tone, or no sound at all (Heikkilä et al., 2015, 2017; Lehmann & Murray, 2005; Moran et al., 2013; Thelen & Murray, 2013; see Matusz et al., 2017 for a review). These studies used old/new memory assessments and found better

recognition memory performance for objects that were experienced audiovisually compared to only visually. However, successful recognition of "old" objects can be accomplished via two mechanisms that are neurally and behaviorally dissociable: familiarity and recollection (Diana et al., 2007; Eichenbaum et al., 2007; Yonelinas, 2002). Familiarity is a general, strength-based type of recognition memory for individual objects, whereas recollection-based recognition includes details of the unique context in which an item was experienced.

In a recent study, we attempted to determine whether multisensory experiences improved memory through familiarity or recollection. We found that the memory benefits of seeing a visual object with a congruent sound were specific to recollection-based recognition and did not extend to familiarity (Duarte et al., 2022). Further, we found better memory for the sounds played during encoding when those sounds were congruent (compared to incongruent or meaningless sounds). This suggests that an object's corresponding sound provided an additional route for accessing the object in memory. However, the results were surprising because recollection typically supports memory for details outside of the studied item itself, including the environment in which an object was encoded or other nearby objects (Mitchell & Johnson, 2009). Thus, one open question is whether the recollection-specific memory benefit would extend to other objects or contexts.

The goal of the current research is to build on our previous findings and investigate the effects of multisensory object encoding on memory for neighboring objects and background context. One possibility is that multisensory encoding could improve memory by increasing within-object feature processing of that object itself at a cost to memory for other objects and the context. Studies have shown this tradeoff between memory for object and context information when a certain object or feature is selectively attended (e.g., Uncapher & Rugg, 2009).

Alternatively, audiovisual objects may promote memory by forming a more elaborate memory representation that includes other objects and contextual details present at encoding. Evidence for this comes from Murray and colleagues (2022) who found that names of faces were better recollected when participants were provided with audiovisual name cues (name tags + spoken names) relative to visual-only cues (written names alone) at encoding. In this case, audiovisual name encoding made it easier to later name the associated face. The primary goal of the present work was therefore to investigate whether, and how, audiovisual object processing affects features of episodic memories beyond the object itself.

Here, we report results from three experiments testing the effect of audiovisual object processing on neighboring visual objects (Experiments 1a & 1b) and source memory for contextual details in which the object was encoded (Experiment 2). In Experiments 1a and 1b, participants viewed audiovisual-visual object pairs or visual-visual object pairs with a control sound, and their memory for each object was tested separately. In Experiment 2, audiovisual objects with semantically congruent or meaningless control sounds were embedded within four different scenes, and source memory for the scenes was tested. To anticipate our results, we found that encoding a congruent audiovisual object did not significantly improve or impair memory for neighboring visual objects. However, congruent audiovisual objects yielded better source memory for the scene context in which they occurred. Our results suggest that multisensory object processing enhances episodic memory, but selectively for source information.

Experiments 1a & 1b

The main objectives of Experiments 1a and 1b were the same: to test the hypothesis that multisensory object processing supports memory for other visual objects present at encoding. In

both experiments (Figure 3.1), participants completed a within-subjects audiovisual encoding task in which two objects were shown on each trial with a sound that was presented centrally (i.e., presented from speakers evenly rather than localized to the right or left). Participants were asked to make a judgment based on the two objects shown. At the end of the experiment, participants were given a surprise recognition memory test of the objects seen during the encoding task. The sound presented on each trial was congruent to one of the two objects (e.g., bicycle + frog + *ribbit*) or neither of them (e.g., bicycle + frog + *white noise*). Therefore, the individual objects were encoded as either an Audiovisual Congruent object (the frog + *ribbit*), a Visual Neighbor (the bicycle neighboring the frog + *ribbit*) or an Audiovisual Control (the frog + *white noise*). The Audiovisual Control condition was chosen instead of a condition with two unisensory visual-only stimuli to control for the possibility that the presence of any sound may bolster memory by enhancing the distinctiveness between objects or by generally increasing attentional alerting. Our Audiovisual Control condition was therefore a stricter test of how sounds impact memory for visual objects.

Experiments 1a and 1b were identical other than the type of judgment made during the encoding task. In Experiment 1a, participants made a size judgment about one of the two objects, indicated by a retroactive cue. In Experiment 1b, participants rated the semantic relatedness of the two objects from 1-6. These encoding tasks differ in that Experiment 1a asks participants to process the two objects individually, but Experiment 1b asks participants to process the two objects in relation to each another. In both Experiments, we used Receiver Operating Characteristics (ROCs) to examine recollection- and familiarity-based recognition memory for each object individually after the encoding task. In this recognition test, participants determined whether a single object had been previously presented by indicating whether each item was old

or new, along with their confidence (Yonelinas et al., 2010). We predicted better recollection for the Visual Neighbor objects than the Audiovisual Control objects in both experiments. However, if the multisensory memory benefit requires relational processing to extend to nearby objects, then we would expect the Visual Neighbors to show improved memory only in Experiment 1b and not Experiment 1a. Together, these experiments provide two methods of investigating the question of whether audiovisual information about a single object impacts recognition memory for a co-occurring neighboring object during encoding: individual processing (Exp. 1a) and relational processing (Exp. 1b).

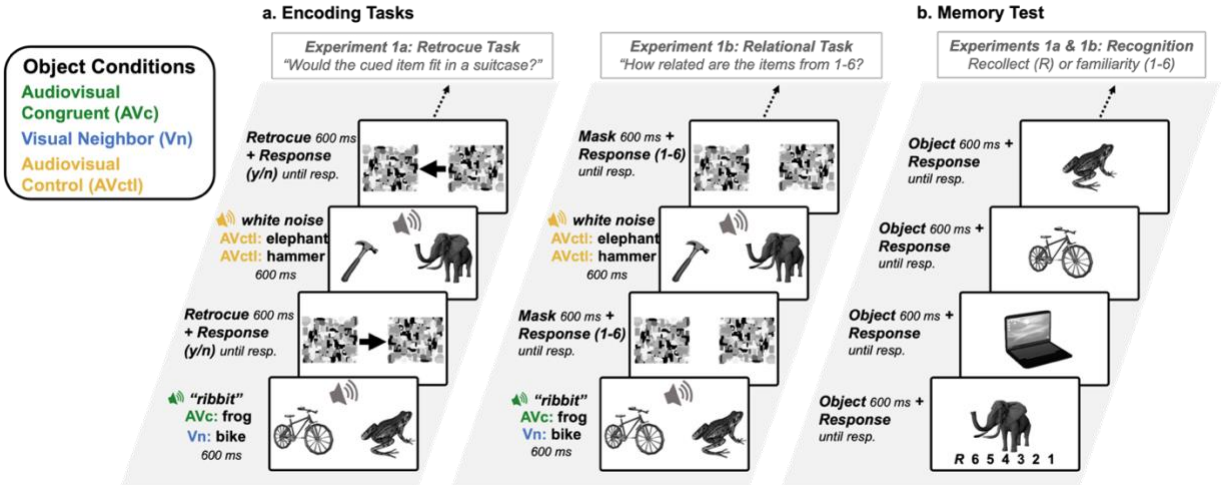


Figure 3.1. Task design for Experiments 1a & 1b. **a.** The Retrocue Encoding Task used in Experiment 1a (left) in which participants make a size judgement about the object in the retroactively cued location, and the Relational Encoding Task used in Experiment 1b (right), in which participants rate (1-6) how related the two items are. **b.** Surprise visual recognition memory test used in both Experiments 1a & 1b, in which participants indicate that they recollect (*R*) the object or indicate recognition confidence by selecting "definitely old," "probably old," "maybe old," "maybe new," "probably new," or "definitely new." Confidence levels are denoted in the figure as 6 (definitely old) to 1 (definitely new).

Method

Participants. Our sample size for Experiments 1a and 1b was determined by an a priori power analysis using the python package Pingouin (Vallat, 2018) with power (1-b) set at 0.95 and $a = 0.05$. Prior data from our lab (Duarte et al., 2022) showed an effect of initial sound

congruency on recognition memory for visual items with an effect size of $\eta^2 = 0.1$, which requires at least 25 participants to detect. To account for potentially poor testing conditions associated with online data collection, we doubled this number to 50, and data were collected in each experiment until we reached this number of participants post-exclusion. The debriefing questionnaire was used to exclude participants by identifying those who participated in a noisy testing environment, exerted little or no effort in completing the study, lacked access to consistent audio (due to glitches, volume changes, or a lack of working speakers), or misunderstood either task (see Supplemental Materials for full list of questions). The sample size and exclusion criteria were pre-registered for Experiment 1a and replicated for Experiment 1b, and can be found on the Open Science Framework (https://osf.io/b3gwt).

Experiment 1a: One hundred nine students from the University of California, Davis, participated in exchange for partial course credit, and 50 were included in analyses (34 identified as female, 13 identified as male, and 3 identified as nonbinary, Mage = 19.02 years; see Supplemental Materials for full sample demographics). Based on our pre-registered exclusion criteria, two participants were excluded from Experiment 1a due to low accuracy the encoding task (below 50%) and 57 were excluded because they were unable to provide examples demonstrating that they understood when to respond with "recollect".

Experiment 1b: Eighty-three students from the University of California, Davis, participated in exchange for partial course credit, and 50 were included in analyses (40 identified as female, 8 identified as male, and 1 identified as nonbinary, Mage = 19.66 years; see Supplemental Materials for full sample demographics). Seven participants were excluded due to low accuracy on the recognition task (below 50%) and 26 were excluded because they were unable to provide examples demonstrating that they understood when to respond with

"recollect". Participants were ineligible to participate in Experiment 1b if they had participated in Experiment 1a.

Materials. Experiments 1a and 1b: A total of 180 images of three-dimensional (3D) models of common objects (e.g., animals, instruments, common household objects; see Supplemental Materials for a full list of items) were gathered from the Unity Asset Store (https://assetstore.unity.com/3d). 3D models were edited in Unity to easy-to-recognize orientations and to reflect the position the objects typically assume when they produce a sound (e.g., the dog model was edited to have an opened mouth, as if it were barking). We used the python package scikit-image (Van der Walt et al., 2014) to remove the image backgrounds, convert them to black and white, and size them to the same dimensions (to fit within 500 x 500 pixels). The real-world sizes of half of the objects in the images were "small" (small enough to fit in a standard suitcase) and the other half were "large." Ninety of the objects were used in the encoding task, and all 180 were used in the recognition memory test. New items in the recognition test were selected from the same categories as the old objects (e.g., animals, instruments, common household objects). The old and new items were not counterbalanced across the encoding and recognition tasks but importantly, the old items, which all had associated sounds, were counterbalanced across the three encoding conditions across participants. Therefore, while overall recognition discrimination between old and new objects may be different, this would not affect the critical comparisons of interest between recognition of items paired with different sounds in the Audiovisual Congruent, Visual Neighbor, or Audiovisual Control encoding conditions. Six different post-stimulus visual masks were manually created using a variety of black, white, and gray geometric shapes arranged in a square the same size as the images.

Natural sounds and white noise sounds were obtained from the Multimost Stimulus Set (Schneider et al., 2008) or found on https://findsounds.com/. Ninety natural sounds corresponded to the items in the encoding task for the audiovisual items, and were selected because the contents were recognizable within the 400 ms duration (see Supplemental Materials for a full list of images and sounds used in each experiment and sound descriptions, see Schneider et al. (2008) for further description of Multimost Stimulus Set stimuli). Fifteen variations of white noise were used for the control condition. All sounds were centrally-presented, 400 ms in duration and amplitude normalized using Audacity (Audacity, 2021). Sounds were kept at 400 ms to match the duration from the Multimost Stimulus Set, though it should be noted that this was shorter than the visual stimulus presentation, which was extended to 600 ms to ensure enough time to view both items presented at once. The sounds and images onset at the same time, so there was 200 ms of visual object presentation after the conclusion of each sound.

Procedure. Experiments 1a and 1b: Participants completed this study online using the Testable platform (https://www.testable.org/). At the start of each experiment, a string of sample beeps was played, and participants were asked to adjust their sound level to a comfortable volume and not to alter it for the remainder of the study. After general instructions, participants completed separate encoding and recognition tasks, followed by a debriefing questionnaire. For Experiment 1a, participants completed the Retrocue Encoding Task, and in Experiment 1b, participants completed the Relational Encoding Task (see below). The Recognition Memory Test and Debriefing Questionnaire were the same in both experiments. Time-to-complete the experiment was not recorded for each participant, though each experiment took approximately 20 minutes to complete during internal pilot testing.

Encoding Tasks. *Experiment 1a: Retrocue Encoding Task.* This consisted of a size judgement task in which participants viewed two items on each trial and received a retroactive cue indicating which item they should base their response on (Figure 3.1a). Forty-five object pairs were presented during this block, for a total of 90 objects (30 Audiovisual Congruent items, 30 Neighboring Visual items, 30 Audiovisual Control items). This resulted in 30 trials with a meaningful congruent sound and 15 trials with a meaningless sound. This was done to ensure there would be an equal number of individual objects in each Object Condition. On each trial, two objects were presented to the left and right of fixation for 600 ms, along with a centrally-presented sound for 400 ms. The sound was semantically congruent with one of the objects or was a meaningless control sound. On the trials with a meaningful sound that matched an item, there was one Audiovisual Congruent object (e.g., frog + *ribbit* in Figure 3.1a) and one Visual Neighbor object (e.g., the bicycle in Figure 3.1a). On trials with a meaningless sound, there were two Audiovisual Control objects (e.g., the hammer and elephant + *white noise* in Figure 3.1a).

After the 600 ms (total) stimulus duration, two post-stimulus visual masks appeared in the position of the two visual images to limit continued visual processing of the objects (Kinsbourne & Warrington, 1962). Between the two visual masks was a retroactive cue (retrocue) in the form of an arrow pointing to the left or the right. Participants were instructed to click on a "yes" or "no" button to indicate whether the item that had been in the cued position would fit in a standard-sized suitcase. The objects were paired such that 15 trials had two small objects, 15 trials had two large objects, and 15 trials had one small and one large object to ensure that participants had to identify both objects on each trial to perform well. Participants were also instructed to ignore the sounds. The masks were presented until the participant responded, and there was a 600 ms inter-trial interval between each trial (not depicted in Figure 3.1a). The items

were counterbalanced across Object Conditions, such that every item appeared as an Audiovisual Congruent item, a Visual Neighbor item, or an Audiovisual Control item across three versions of the experiment between participants. The probability of the retrocue pointing to the left or to the right was equal across all trial types and Object Conditions. The retrocue/size judgement task was designed to require participants to identify both visual items during the initial presentation, without specifically freeing them to focus on either individual item and to prevent participants from anticipating a memory test, making the subsequent recognition block a test of incidental object memory.

Experiment 1b: Relational Encoding Task. The encoding task for Experiment 1b consisted of an item pair relation judgement task in which participants view two items on each trial and rate how likely the two objects are to be seen together in the real world from 1 (rarely/never) to 6 (often/always) (Figure 3.1a, right). The encoding block was similar to Experiment 1 in that there were 45 visual object pairs presented during the block, for a total of 90 visual items (30 Audiovisual Congruent items, 30 Neighboring Visual items, 30 Audiovisual Control items). The items in this task were paired such that half of the items were likely to be seen together in the world often or always (e.g., goat & pig), and the other half were likely to be seen together rarely or never (e.g., penguin & microwave). As with Experiment 1, on each trial, two visual items were presented to the left and right of fixation for 600 ms, along with a centrally-presented sound for 400 ms. The sound was semantically congruent to one of the visual items, or was a control white noise sound. After the 600 ms stimulus, two post-stimulus visual masks appeared in the position of the two visual images, which functioned to limit continued visual processing of the object in order to accentuate the timing co-occurrence of the visual and auditory stimuli (Kinsbourne & Warrington, 1962). During this time, participants indicated, from

1 to 6, how likely these items were to be seen together in the real world. As in Experiment 1a, participants were instructed to ignore the sounds. The masks were presented until the participant responded, and there was a 600 ms inter-trial interval between each trial (not depicted in Figure 3.1a). As in Experiment 1a, the items were counterbalanced across Object Conditions, such that every item appeared as an Audiovisual Congruent item, a Visual Neighbor item, or an Audiovisual Control item across three versions of the experiment between participants. This task was designed to encourage participants to not only identify both items individually, but to consider their relation to one another to facilitate between-object binding, and to prevent participants from anticipating a memory test.

Recognition Memory Test. *Experiments 1a and 1b:* This memory test was designed to dissociate between recollection- and familiarity-based recognition in accordance with the dual-process signal detection model using responses to six response criteria to construct receiver operating characteristics (ROCs; Yonelinas, 1994). However, there are multiple methods for dissociating between recollection- and familiarity- based recognition, which differ in their assumptions of the characteristics of each process (see Yonelinas et al., 2010). We therefore included a subjective, introspective measure of recollection to allow us to conduct a supplemental analysis using the remember/know (recollect/familiar) procedure to assess whether results converge across multiple process-dissociation methods (Tulving, 1985; see Data Analysis, Appendix A).

In the visual-only memory test, the 90 old images were intermixed with 90 new images for a total of 180 trials. On each trial, a single visual stimulus was presented for 600 ms, and participants could respond by clicking on buttons corresponding to an introspective report of "recollect" or to one of six other response criteria: "definitely old," "probably old," "maybe old,"

"maybe new," "probably new," or "definitely new" (Figure 3.1b). Participant instructions included a description and example of the difference between a "recollect" response and any "old" response, explaining that "recollect" should only be pressed if the participant was sure that they had seen the item before *and* they could recollect some qualitative information about the encoding event, such as their feelings about the item or what they thought about when they initially saw it. The item presented on each trial corresponded to one of the three Object Conditions (Audiovisual Congruent, Visual Neighbor, Audiovisual Control) from the encoding task. The order of objects was randomized for each participant.

Debriefing Questionnaire. *Experiments 1a and 1b*: After the experiment, participants responded to questions on a debriefing questionnaire, which allowed us to assess the quality of the testing environment and stimulus presentation. This survey included questions about the testing environment, the subjective volume and quality of the auditory stimuli, whether the volume was adjusted during the experiment, whether any glitches or lags between audiovisual stimuli were experienced, among others (see Supplemental Materials Table 3.4 for a full list of questions). As this experiment was completed remotely, responses to this survey were used to exclude participants when the testing environment or stimulus presentations were not of adequate quality. To ensure that participants understood the recognition task, the debriefing survey included a question asking whether participants understood when they were supposed to press the "recollect" button. We also included a free-response question asking for an example of information they used to judge an item as recollected rather than definitely old on one of the trials. While this question only asked for a single example from each participant, we performed exploratory analyses on these responses to assess the types of details participants found salient enough to mention.

Data Analysis. The design, hypotheses, and statistical analyses for Experiment 1a were pre-registered prior to data collection on the Open Science Framework (https://osf.io/b3gwt), and the same statistical analysis approach was used for Experiment 1b.

Experiments 1a and 1b: We compared overall memory performance for items between Object Conditions using the observed area under the curve (AUC). To directly assess effects of Object Conditions on recollection- and familiarity-based recognition, we fit the Dual-Process Signal Detection (DPSD) model to our confidence data (Yonelinas, 1994). Subjective "recollect" responses were used to perform supplementary remember/know analyses to assess the convergence of our results across multiple recollection and familiarity process-dissociation methods (see Supplemental Materials Appendix A). We also performed an exploratory analysis of responses to the open-ended debriefing survey prompts asking participants to report an example of information they used to base their "recollect" responses on. Additionally, we have included mean hit rate (% correct recognition of old items) and false alarm rate (% incorrect recognition of new items) across response criteria for the recognition tasks in the Table 3.1. Performance on the encoding tasks for both Experiments (% correct) are reported in Table 3.7 of the Supplemental Materials, and exploratory analyses of the relation between encoding and memory performance are reported below. Raw data files for both experiments are publicly available on the Open Science Framework (https://osf.io/sep3r/).

ROC Analysis. Experiments 1a and 1b: For the ROC analysis, we calculated the cumulative hit rates (the proportion of old items correctly identified as old) for items in each Object Condition at each response criterion, and calculated the false alarm rates (the proportion of new items incorrectly identified as old) at each response criterion to analyze the underlying ROC (Yonelinas & Parks, 2007). Each subsequent point on an ROC curve relates the hit and

false alarm rates as participants increasingly relax their criteria for classifying an item as *old*, from "definitely old" to "definitely new." In line with the dual-process signal detection (DPSD) analysis of ROCs, the leftmost point includes both "recollect" and "definitely old" responses (Yonelinas et al., 2010). While the remember-know procedure relies on subjective reports of recollection- versus familiarity-based recognition to dissociate these processes, the DPSD model instead estimates these processes based on hit rates and false alarm rates at each level of confidence through ROCs. The benefit of this ROC method is that it does not require participants to accurately introspect and assess the source of their own memory. However, we included the "recollect" response option to allow us to conduct remember-know analyses to assess whether our results converge across different commonly-used process-dissociation methods. We have included remember/know analyses for both Experiments 1a and 1b in the Supplemental Materials, and we note here that all results converge with the ROC results, except for the specific case noted in the Experiment 1 Discussion and Supplement (Appendix A).

For statistical analyses, individual ROCs were constructed for each participant at each level of Object Condition, and the points in the ROC graph of Figure 3.2a reflect the average observed hit and false alarm rates for these conditions across participants. The same false alarms rates for new items were used across the three Object Conditions, which allows us to identify each participants' individual criteria for judging an item as new across confidence levels to compare to hit rates at the same confidence levels. DPSD models were fit to each participants' ROCs, and the average ROC model for each group is shown in Figure 3.2a. To compare overall differences in memory performance between conditions, AUC was calculated using the trapezoidal rule for each participant's observed ROCs in each Object Condition and compared via one-way RM ANOVA and Bonferroni corrected post-hoc pairwise t-tests. Parameter

estimates derived from DPSD model-based ROCs were used to compare two constructs of interest from the dual-process model of recognition memory, namely the *y-intercept*, which estimates recollection, and *d-prime* (*d'*), which estimates familiarity. In the DPSD model, the y-intercept estimates the hit rate when the false alarm rate is equal to 0, making it a threshold measure of memory that represents recollection. Model-derived d' measures hit rates relative to false alarm rates across the entirety of the curve, which quantifies the contribution of familiarity. These estimates were also compared via individual one-way RM ANOVAs and Bonferroni adjusted post-hoc pairwise t-tests.

Finally, to assess whether encoding task performance is related to recognition memory performance, we ran exploratory correlation analyses for both experiments. For both experiments, encoding task performance was assessed as percent correct, and this was related to hit rate on the recognition tasks.

Debriefing Questionnaire Analysis. Experiments 1a and 1b: To perform an exploratory assessment of the types of details that were retrieved about objects on trials for which participants respond with "recollect," we coded the open-ended responses to the debriefing survey for mentions of specific items and features recollected from the encoding task (Figure 3.2c). All responses included mention of an item from the encoding task along with a feature that was recollected. Responses that referred to objects and their accompanying sound were labeled as "Sound" recollections (e.g., "I remembered the dog because it was shown along with a 'bark' sound"). Responses that referred to the item that an object was paired with at encoding were labeled as "Paired Item" (e.g., "I remembered the toaster because it was paired with the tea kettle"). Other recollected aspects of the encoding experience related to the task or personal

opinions or thoughts were labeled as “Other” recollections (e.g., “I remembered the elephant because elephants are my mom’s favorite animal”).

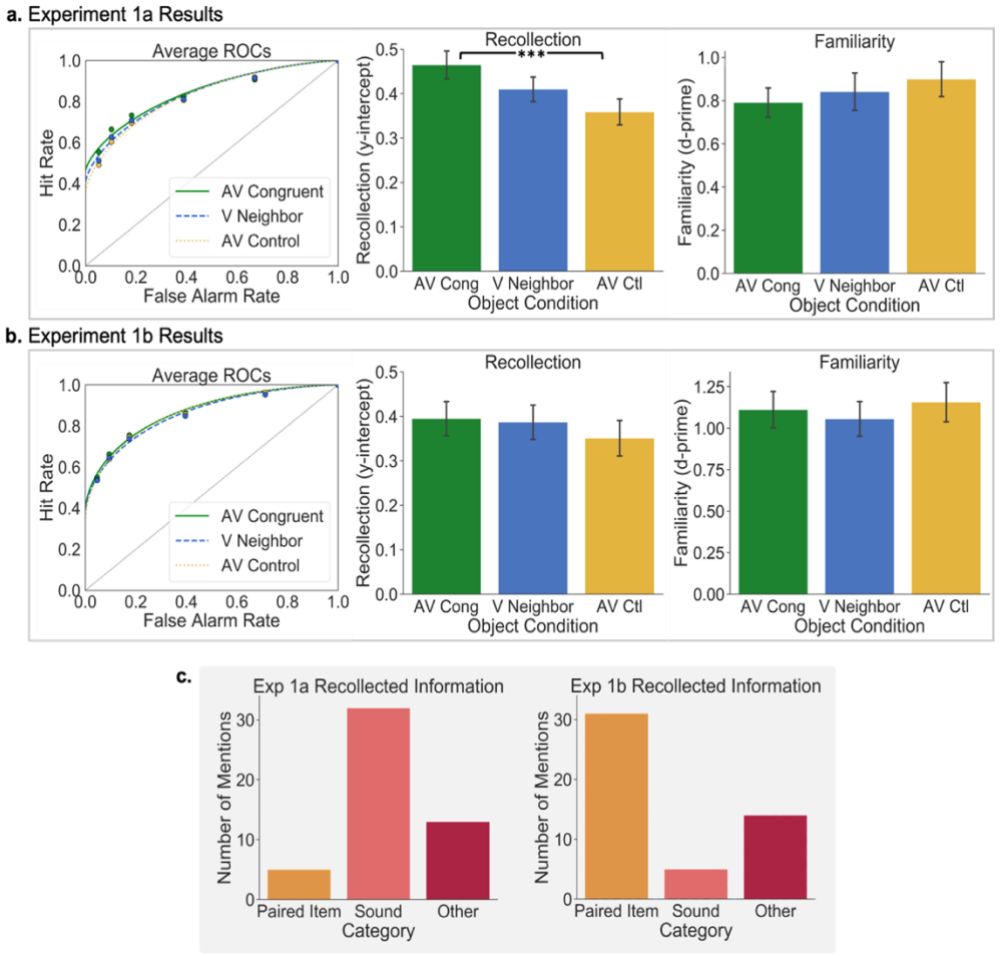


Figure 3.2. Results for Experiments 1a and 1b. **a.** The left graph depicts the averaged observed ROCs (points) for each Object Condition from Experiment 1a and corresponding Dual-Process Signal Detection (DPSD) functions. The center and right graphs show mean *y-intercepts* and *d-prime* as DPSD-based metrics of recollection and familiarity, respectively. The center graph shows recollection is greater for items encoded in the Audiovisual Congruent (AV Cong) condition than the Audiovisual Control (AV Ctl) condition. **b.** The left graph depicts ROCs for Experiment 1b, and the center and right graphs show mean *y-intercepts* and *d-prime* as DPSD-based metrics of recollection and familiarity, respectively. **c.** Mentions of each type of recollected information from the debriefing questionnaire, including mentions of the item the object was paired with during encoding (Paired Item), the sound played at encoding (Sound), or any other information about the task or objects (Other). All error bars illustrate standard error of the mean.

Table 3.1. Mean hit rates (% correct recognition of old items) and false alarm rates (% incorrect recognition of new items) for the recognition tasks, and mean DPSD model parameters for overall recognition memory (*AUC*), recollection (*y-intercept*), and familiarity (*d’*) for items in each Object Condition for Experiments 1a and 1b.

				DPSD Model Parameters		
Exp.	Object Condition	Hit Rate	False Alarm	*AUC*	*y-intercept*	*d’*

1a	Audiovisual Congruent	73.33(13.50)	18.22(10.27)	0.80(0.09)	0.46(0.22)	0.79(0.48)
	Visual Neighbor	70.80(12.43)	18.22(10.27)	0.80(0.09)	0.41(0.20)	0.84(0.57)
	Audiovisual Control	69.67(12.29)	18.22(10.27)	0.79(0.08)	0.36(0.21)	0.84(0.61)
1b	Audiovisual Congruent	75.53(13.84)	17.42(13.98)	0.82(0.12)	0.39(0.27)	1.11(0.78)
	Visual Neighbor	73.73(14.02)	17.42(13.98)	0.81(0.12)	0.39(0.27)	1.06(0.74)
	Audiovisual Control	74.60(15.24)	17.42(13.98)	0.82(0.12)	0.35(0.28)	1.16(0.84)

Standard deviations are shown in parentheses.

Results

ROC Analysis.

Recollection and Familiarity. ROC analyses were conducted in order to determine if recollection or familiarity for objects differed as a function of their occurrence with a congruent sound, appearance in the context of the other object with a congruent sound, or in the presence of white noise.

Experiment 1a: A one-way repeated measures analysis of variance (RM ANOVA) showed a significant effect of Object Condition on recollection (*y-intercept*), $F(2, 98) = 6.55$, $p = 0.002$, $\eta_p^2 = 0.12$ (Figure 3.2), with higher y-intercepts for items in the Audiovisual Congruent condition than in the Audiovisual Control condition $t(49) = 4.41$, $p = 0.0002$, *Cohen's d* = 0.49; there was no significant difference between items in the Audiovisual Congruent and Neighboring Visual conditions, $t(49) = 1.71$, $p = 0.28$, *Cohen's d* = 0.26, or the Audiovisual Control and Neighboring Visual conditions, $t(49) = -1.64$, $p = 0.32$, *Cohen's d* = -0.25. Bayes Factors provided very strong evidence for the difference between y-intercepts for the Audiovisual Congruent and Audiovisual Control conditions ($BF_{10} = 386.31$), and only anecdotal evidence for the null hypothesis for differences between y-intercepts for the Audiovisual Congruent and Neighboring Visual conditions ($BF_{01} = 1.67$) and Audiovisual Control and Neighboring Visual conditions ($BF_{01} = 1.85$). A one-way RM ANOVA did not show a significant effect of Object Condition on familiarity (*d'*), $F(2, 98) = 0.66$, $p = 0.52$, $\eta_p^2 = 0.01$ (Figure 3.2a). These results

indicate that recollection was higher for the audiovisual congruent object than the other objects, replicating our previous results. However, there was no memory benefit for the Neighboring Visual object that co-occurred with the Audiovisual Congruent object, suggesting that the benefit conferred by the congruent multisensory event did not transfer to other objects in the same visual context.

Experiment 1b: A one-way RM ANOVA showed no significant effect of Object Condition on recollection (*y-intercept*), $F(2, 98) = 0.59$, $p = 0.56$, $\eta_p^2 = 0.012$, or familiarity (*d-prime*), $F(2, 98) = 0.51$, $p = 0.60$, $\eta_p^2 = 0.010$. (Figure 3.2b). These results suggest that the auditory stimulus had no effect on visual memory when the encoding task explicitly required a relational comparison of the two visual objects.

AUC. Experiment 1a: A one-way RM ANOVA did not show a significant effect of Object Condition on general recognition memory (AUC) $F(2, 98) = 0.65$, $p = 0.53$, $\eta_p^2 = 0.01$. This suggests that the auditory stimulus did not impact recognition memory accuracy beyond recollection-based recognition, which is unsurprising given that familiarity was also not significantly impacted.

Experiment 1b: A one-way RM ANOVA did not show a significant effect of Object Condition on general recognition memory (AUC) $F(2, 98) = 1.06$, $p = 0.35$, $\eta_p^2 = 0.021$. This suggests that the sound stimuli did not have an impact on overall recognition memory.

Encoding & Memory Performance Relationship. For both experiments, we explored whether there were associations between successful encoding task performance and subsequent recognition.

Experiment 1a: The correlation was not significant between performance on the encoding task (M = 60.27%, SD = 5.60%) and the hit rate on the recognition task across participants (M = 76.52%, SD = 6.76 %; $r(50) = 0.009$, $p = 0.9504$).

Experiment 1b: There was a significant positive correlation between performance on the encoding task (M = 75.93%, SD = 18.29 %) and the hit rate on the recognition task (M = 78.60%, SD = 10.75%; $r(50) = 0.41$, $p = 0.003$).

Debriefing Questionnaire. We analyzed participants' responses to the open-ended question in the debriefing questionnaire, which asked them to describe the features of objects they relied on to provide their recollection responses.

Experiment 1a: Out of a total of 50 responses, five (10%) mentioned the Paired Item, 32 (64%) mentioned the Sound, and 13 (26%) mentioned some Other aspect of the encoded event (Figure 3.2c). This suggests that out of the features present at encoding that could have been recollected (e.g., the paired item, the sound, task-related information), the sounds matching Audiovisual Congruent stimuli were recollected and stood out compared to other features of the encoded event.

Experiment 1b: Out of a total of 50 responses, 31 (62%) mentioned the Paired Item, five (10%) mentioned the Sound, and 14 (28%) mentioned some Other aspect of the encoded event (Figure 3.2c). This suggests that out of the features present at encoding that could have been recollected (e.g., the paired item, the sound, task-related information), participants were more likely to mention the Paired Item later on than the sound. This is in contrast to Experiment 1a (Figure 3.2c) which showed the opposite pattern for the Paired Item and Sound mentions.

Experiment 1a & 1b Discussion

The goal of these complementary experiments was to assess whether multisensory object processing supports memory for nearby visual objects present at encoding using two different encoding tasks. In the encoding tasks, participants made size judgements about one of the two objects presented on each trial (Exp. 1a) or rated their semantic relatedness (Exp. 1b), in tasks that were designed to assess the impacts of multisensory processing on memory for nearby objects under conditions that elicit more individual versus relational processing, respectively. The ROC analysis for Experiment 1a replicated our previous findings, indicating that recollection-based recognition memory was superior for objects paired with congruent sounds during encoding, while familiarity was not different between Object Conditions. However, results did not align with our main hypothesis that recollection would also be significantly better for Neighboring Visual objects than Audiovisual Control objects. Bayes Factors suggest anecdotal evidence that this is the case, though this comparison did not reach statistical significance. The remember-know analysis (see Appendix A) did actually show that Visual Neighbor items were remembered (recollected) at significantly higher rates than the Audiovisual Control objects. Unlike the ROC analysis, the remember-know procedure relies on participants' ability to accurately determine the source of their own memory. This provides some evidence of a memory benefit for visual objects encoded within the context of congruent audiovisual objects. Overall, results from Experiment 1a did not provide strong, converging evidence that the presence of audiovisual objects also supports individual memory for a neighboring visual item. However, they did replicate our previous findings showing improved memory for the congruent audiovisual object itself and that the recollection benefit did not come at a cost to memory for nearby visual objects.

Experiment 1b was motivated by the possibility that the retrocue task in Experiment 1a emphasized processing each item alone, and that this may have led objects to be processed *more* individually than they would be in a naturalistic setting where they might be found co-occurring. However, results from Experiment 1b did not align with our initial hypotheses or with results of Experiment 1a. When the encoding task emphasized the associative relations between objects, the Audiovisual Congruent objects were no longer better recollected than the Audiovisual Control or Neighboring Visual objects. This may be because elaboration on the meaning and relation between the two objects improved memory over-and-above the memory benefits of audiovisual encoding. This idea is supported by the overall higher hit rates and lower false alarm rates compared to Experiment 1a (see Table 3.1). Further, in Experiment 1a, Debriefing survey responses showed that participants most often mentioned congruent sounds when making recollection judgements rather than the paired item or some other detail from the encoding task, whereas in Experiment 1b, there were many more mentions of the paired objects than sounds. This could mean that the sounds were the most notable features aiding recollection in Experiment 1a, whereas the paired items drove recollection more in Experiment 1b. While the lack of effect of congruent audiovisual object processing on memory was unexpected, this provides new insight on the conditions under which audiovisual processing does not facilitate memory. However, more research is necessary to fully understand these boundary conditions in the multisensory benefit.

Contrary to expectations, neither Experiment provided strong evidence that the audiovisual recollection benefit for a single object extends to nearby unisensory objects. One reason for this may be that sounds only enhance encoding of the object producing the sound and do not extend to multi-object memories. Another reason, however, may be that our recognition

memory task only measured memory for single objects and not for the association between the objects. The multisensory event may have generated a benefit for associative memory between the two objects, but our memory test was not sensitive to benefits in associative learning given that we tested memory for objects individually. This idea motivated Experiment 2, in which we test whether encoding audiovisual objects improves associative source memory for the context in which it was encoded.

Experiment 2

The goal of this experiment was to test the hypothesis that multisensory object encoding facilitates memory for the context in which an object is embedded. Unlike the previous experiments, the context was defined by visual scenes within which the target objects were embedded. This approach emulates more naturalistic associations between objects and their context. To examine memory for the context, we assessed source memory rather than individual item recognition memory. Participants first familiarized themselves with four virtual environments: a barn area, a pond area, a living room, and a bedroom, learning their layouts and the objects in them (Figure 3.3a, b). They then completed an audiovisual encoding task, in which they were presented with target objects embedded within scenes from these environments, along with a sound that was congruent to the target object, or a control sound (white noise). Participants searched for the target object (demarcated with an X) and made a size judgement about that object (Figure 3.3c). In a surprise source memory test, participants indicated which of the four environments the item had been encoded in, their confidence in that memory, and, on a map of that environment, clicked on where the object had been located (Figure 3.3d).

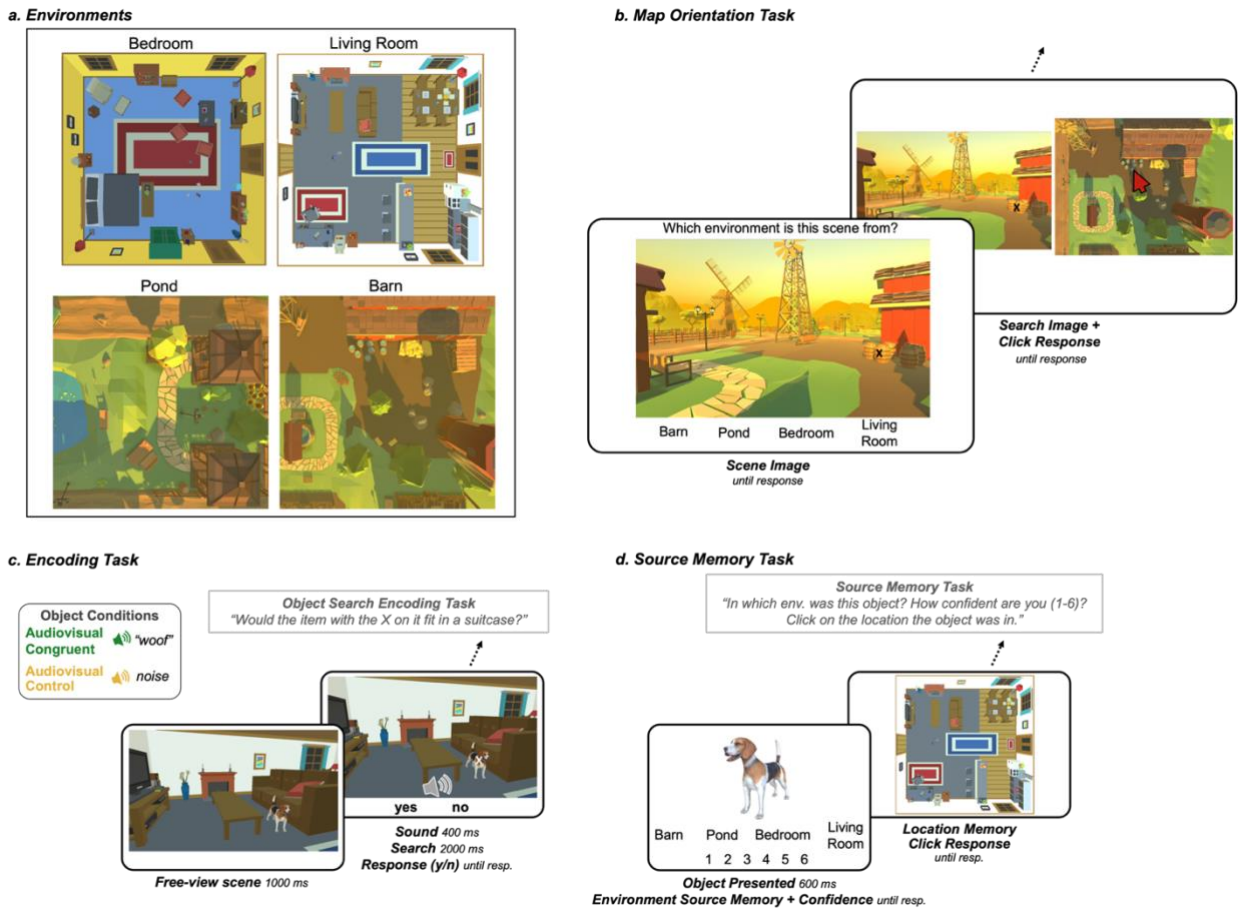


Figure 3.3. Experiment 2 design. **a.** Overhead view of the four environments used in this experiment, two indoor house environments, and two outdoor farm environments. **b.** Map Orientation Task in which participants identify which environment a scene image is from, then search for a target "x" in the scene and click on the location on the map where the x is located. **c.** Object Search Encoding Task in which participants search for the object with an "x" on it and make a size judgement about that object. **d.** Source Memory Task in which participants indicate in which environment they encoded the object (Environment Source Memory), confidence (1-6) in that judgement (Confidence), and then click on the location on the map where the object was encoded (Location Memory).

Method

Participants. 209 students from the University of California, Davis, participated in exchange for partial course credit, and 68 students were included in analyses (53 identified as female, 12 identified as male, and 3 identified as nonbinary, Mage = 19.40 years; see Supplement Table 3.6 for the full sample demographics). Our pre-registered exclusion criteria were based on performance (less than chance level performance on either the encoding (n = 6) or

memory tasks (n = 1)), use of the entire range of confidence options on the memory test (n = 112), and based on responses to the debriefing survey. From the debriefing survey, participants were excluded if they participated in a somewhat or very noisy testing environment, did not have consistent audio available during the study (n = 17), exerted no effort or little effort in completing the study (n = 5), or they reported that they did not understand any of the tasks (n = 0). An a-priori power analysis for one-tailed t-tests determined that we would need a sample size of 34, and as with Experiments 1a and 1b, we doubled this for our target sample of 68. This sample size was pre-registered, and data were collected until we reached 68 participants post-exclusion. Participants were ineligible to participate in Experiment 2 if they had participated in Experiments 1a or 1b.

Materials. Materials for this experiment included most of the same objects as the previous experiments, embedded within scenes from virtual environments, and deviations are described below.

***Environments*.** Four environments were obtained from the Unity 3D Asset store (https://assetstore.unity.com/3d). These 3D environments consisted of two house environments (Bedroom & Living Room) and two farm environments (Barn & Pond; Figure 3.3a). These environments were chosen because the majority of the objects used in the previous experiments reasonably fit within either the farm or house environments. Further, having an indoor set of environments and an outdoor set of environments allowed us to counterbalance the objects across two environments per object to control for memory biases based on the memorability of features of a specific environment, rather than the memorability of objects based on the Object Condition. The two farm environments were each approximately 7 x 7 meters, and the house environments were each approximately 5 x 5 meters. The environments were each filled with

items typical to that space (e.g., the barn area has hay bales, a wheelbarrow, barrels). The overhead view of the map of each environment was split into a 5 x 5 grid of possible locations for objects within the encoding task. Of the 25 possible locations per environment, 3 were removed from each because they did not contain a "valid" placement for objects (e.g., in the pond). This grid structure was implemented to ensure that visual objects were evenly distributed across the environments. Each cell was edited to include at least one object in order to provide nearby contextual items for each encoded object. For the *Environment Orientation Task* (described below), point-of-view images of the environments were taken from ground level in various areas around the map without additional experimental objects added in. Each map also had map images taken from above, which were used in the map orientation task and the memory test.

Objects. We used 88 visual objects, which were divided into two groups, with 44 corresponding to the farm environments, and the other 44 corresponding to the house environments (see Supplemental Materials for full lists of items). The majority of these objects were the same as those used in Experiments 1a and 1b, though in order to more closely match each environment, six objects were replaced with four new ones which better matched the chosen environments. Additionally, some items (e.g., the leopard) were now used as *toys* within the environments rather than as the real version of the objects, which further served to improve the correspondence of objects and their environments. New visual objects also came from the Unity 3D asset store, and new sounds that were added came from the Multimost Stimulus Set. In previous studies, we presented objects in grayscale, though here, objects and scenes were presented in color in order to increase the realism of the environments and objects. During the encoding phase, objects were presented within the environments in scenes (see below for

details), and during the memory test, objects were shown on their own, edited to fit within 500 x 500 pixels. Sounds used in this experiment were either congruent to the object or a control white noise sound, which were the same control sounds used in the previous experiments.

***Scenes*.** Scenes were constructed by placing objects in one of the grid cells within the 5 x 5 map of locations excluding unrealistic locations for placing our objects (e.g., the pond). This left 22 locations for target object per environment. The objects in each cell were randomized so that performance in the memory phase could not be based on guessing related to typical locations of items. Therefore, some objects were in atypical locations within the environments (e.g., a hammer on a bench), but none were so out of place that they would be considered impossible. For each scene image, objects were placed anywhere within their designated cell, and screen captures were taken with the objects at ground-level and within the central-third of the image. The objects were a variety of sizes to emulate naturalistic size variation. Each object only appeared once, and was not visible in the background of other images, though the rest of the background images for each environment was consistent across the scenes. Each object was counterbalanced across Object Conditions and its two corresponding Environments, such that each scene could be paired with a congruent sound or with a control sound, and each object appeared in both environments once between subjects. For example, across subjects, the sheep appeared in both the barn area and in the pond area. For the target search phase of each trial, black or white Xs were placed on the target object (depending on which was more easily visible) on the target object. Each image was presented at 1,000 by 1,000 pixels.

Procedure. As with the previous experiments, this experiment began with sample beeps for participants to use for sound level calibration, and all stimuli were presented online via personal computers through the online stimulus presentation software Testable

(https://www.testable.org/). Participants completed the *Map Familiarization Phase*, then the *Map Orientation Task*, followed by the *Object Search Encoding Task*, then the surprise *Source Memory Task*, and finally the *Debriefing Questionnaire*. Time-to-complete the experiment was not recorded for each participant, but the experiment took approximately 30 minutes to complete during internal pilot testing.

***Map Familiarization Phase*.** This block of stimuli was used to introduce participants to the layout of each environment and information within them. The goal of this was to encourage participants to consider features of the environmental context during the encoding task, even though there were no explicit instructions to do so during that task. In this block, participants were shown successive images of each environment, including an overhead map of the environment and four images of each environment from first-person perspective (Supplemental Materials Figure 3.2 for examples). They were told to acquaint themselves with the environments because they would be asked questions about them in the next task. Participants were first shown the map and a broad view of an environment, along with a short description of what the environment contains (e.g., "This is The Barn. It contains objects like a barn, a well, hay, and barrels. Have a look around"). They were then shown three additional pictures of each environment from a first-person perspective, followed by all five images together. This was done for each of the four environments, and the order of the environments was randomized for each participant. Each trial was displayed for five seconds before allowing participants to advance at their own pace.

***Map orientation task*.** The purpose of this self-paced task was for participants to practice identifying which environment a given scene came from, and matching the first-person perspective scenes to their locations on the overhead map of the environments in order to

complete the memory task later on. On each trial, participants viewed a scene from one of the four environments and indicated which environment it came from by clicking an on-screen button with the environment labels (Figure 3.3b). They were then shown the image next to the map of the correct environment. On these trials, a small target (an X) was added to an object or location on the scene, and participants were asked to click on the overhead map the location that the target was shown in the scene. There were six scenes randomly presented from each environment, for a total of 24 environment identification trials, and 24 localization trials, and 48 trials in this block overall.

The encoding task consisted of 88 trials of a visual search and size judgement task within scenes from the environments. On each trial, participants viewed a scene from one of the environments for one second before the onset of a target X on top of the target object and a sound. The sound corresponded to one of the two Object Conditions, such that it was congruent with the target object (Audiovisual Congruent), or it was composed of white noise (Audiovisual Control; Figure 3.3c). After the onset of the target, the image disappeared after 2000 ms, and participants were asked to indicate whether the item with the target X on it would fit in a standard suitcase. Once participants responded, there was a 600 ms inter-trial interval. Scenes from each environment were presented randomly. Across versions of the experiment, the Object Conditions were counterbalanced, such that every object was paired with a congruent sound and a control sound in one version of the experiment. Additionally, each target item was counterbalanced across its two congruent environments (e.g., the stapler was shown in both the living room and bedroom). This counterbalancing scheme resulted in 8 versions of the experiment, to which participants were pseudorandomly assigned by Testable to result in an even distribution of experiment versions completed across participants. This task was designed to

encourage participants to process the scenes prior to searching for and making a size judgment about the target object which was either an audiovisual or control object. This allowed us to later assess the impact of the Object Conditions on memory for the target objects and the contexts in which they appeared.

Source Memory Test. The surprise source memory test included 88 two-part trials to test memory for the environments in which each object was encoded (Environment Source Memory), confidence in that judgment (Confidence), and memory for the precise location within the environment that the item was originally placed (Location Memory). Environment Source Memory: On each self-paced trial, participants saw a target object from the encoding task and indicated which of the four environments, "barn," "pond," "bedroom," and "living room", the object was originally located in (Figure 3.3c). Confidence: Participants indicated their confidence in their Source Memory judgement on a scale from 1 (not confident at all) to 6 (extremely confident). Location Memory: Participants were shown a map of the environment that they selected, and asked to use their mouse or trackpad to click on the location within the map where the object originally appeared. If the participant selected the wrong environment during the first part of the trial, they were taken to the incorrect map, and only the Environment Source Memory of these trials were analyzed. This task was designed to test whether Object Condition influenced memory for the context in which an object was encoded, confidence in that memory, and precision of the visuospatial memory for that item within the environment. Participants were instructed to use the entire range of response options across trials.

Debriefing Questionnaire. As with our previous experiments, we had participants complete a debriefing survey after the end of the experiment. The survey was similar to previous experiments, with the addition of questions specific to this study, such as whether participants

expected a memory test, whether there were any individual tasks of the study in which they put in less effort, and an introspective question asking for participants to give an example of what they believe helped them remember the items that they did remember (see Supplemental Materials for a full list of questions).

Data Analysis. Our pre-registered analyses included comparisons between source memory performance for items paired with congruent sounds and control sounds during the Object Search Encoding Task. We deviated from some aspects of the pre-registered analysis plan to opt for a more appropriate analysis for our data, as described below, though the hypotheses tested remain the same. The raw data and results from pre-registered analyses can be found in the experiment folder on the OSF page (https://osf.io/sep3r/) along with the analyses presented in the paper. Results from the Map Orientation Task were used for comparison with Location Memory and descriptive statistics for that task are reported in the Location Memory section of the results.

***Source Memory*.** Environment Source Memory: To assess source memory for the environment in which objects were encoded across Object Conditions, we performed a t-test comparing accuracy (*percent correct*) between the Audiovisual Congruent and Audiovisual Control Object Conditions. We compared percent correct rather than d-prime because percent correct is a more common measure for analysis of *m*-AFC tasks and because both of these measures convey relative performance accurately within forced-choice tasks (Brady et al., 2021; Stanislaw & Todorov, 1999).

Confidence: As an additional measure of memory strength, we compared confidence ratings between conditions. To do this, we took the average confidence rating (1-6) for each participant in each condition for correct responses (hits), and compared these averages between conditions using a t-test.

Although our pre-registered analysis plan included an ROC analysis based on confidence responses, hit rates, and false alarm rates, we ultimately decided the number of items that we had per condition would not be enough to properly characterize ROC functions. Yonelinas & Parks (2007) state that ROCs constructed from fewer than 50 responses per condition (i.e., 50 old items and 50 new items) can be noisy and irregularly shaped, impairing the ability to properly characterize the function. Whereas our Experiments 1a and 1b had 30 items per condition and 90 new items (total of 180 items), Experiment 2 only had 45 items per condition, all of which had been seen before. We deemed this to be too few trials to reliably carry out the ROC analysis. However, because confidence ratings have been shown to track with accuracy and memory strength (Brady et al., 2021; Mickes et al., 2011), we have included the confidence data here using the alternative analysis plan.

Location Memory: Accuracy (Euclidean distance) was compared between Object Conditions using a paired-samples t-test. Trials with incorrect Environment Source Memory were excluded from this analysis.

Encoding & Memory Performance Relationship. To assess whether relationships exist between encoding task performance and recognition memory performance, we ran an exploratory Pearson correlation analysis between percent correct on the size judgement encoding task, compared to percent correct on the environment judgement of the memory test.

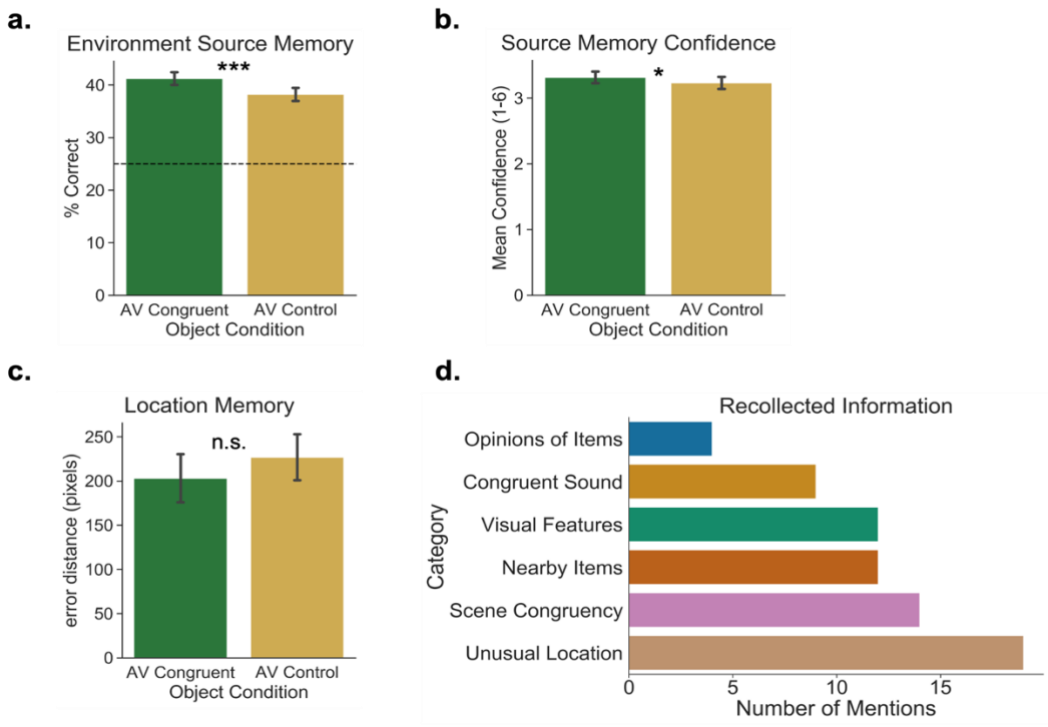


Figure 3.4. Experiment 2 results. **a.** Average accuracy (% correct) for environmental context source memory between Object Conditions. Accuracy was significantly higher in the Audiovisual Congruent condition than the Audiovisual Control condition. **b.** Source memory confidence between Object Conditions for correct environmental context source memory. Confidence was significantly higher for items in the Audiovisual Congruent condition than those in the Audiovisual Control condition. **c.** Error of Location Memory for where the object was located in the scene. **d.** Mentions of information that was reported to facilitate memory for the environment the object was encoded in as reported in the debriefing questionnaire. Error bars illustrate standard error of the mean.

Debriefing Questionnaire. We performed an exploratory analysis of responses to the open-ended debriefing survey prompt asking participants to report an example of something they believed helped them remember the environment that objects were located in during encoding. For this analysis, we coded responses for mentions of different types of information (e.g., mentions of memory for sounds, visual features, or background scene objects) to assess themes across participants. Responses that included mentions of multiple types of information were counted in each of those categories, so the total number of responses exceeds the number of participants. We report counts of mentions of information from the emergent categories in Figure 3.4.

Results

Source Memory.

Environment Source Memory. For this analysis, we assessed memory for the environment in which each item was encoded across the two Object Conditions. A one-tailed, paired-samples t-test revealed a significant effect of Object Condition on Environment Source Memory (percent correct), $t(67) = 2.47$, $p = 0.008$, *Cohen's d* = 0.30, such that memory performance was better for objects encoded in the Audiovisual Congruent condition ($M = 0.41$, $SD = 0.10$) than for objects encoded in the Audiovisual Control condition ($M = 0.38$, $SD = 0.10$) (Figure 3.4a). Additionally, the Bayes Factor suggests moderate evidence for this finding ($BF_{10} = 4.47$). Our pre-registered t-test comparing *d-prime* rather than percent correct between the two Object Conditions yielded an effect with almost identical statistics, which are reported in the supplemental materials on OSF (https://osf.io/sep3r/).

To visualize performance across the environments, we plotted a confusion matrix with the true encoding environments and the response environments (Figure 3.5). False alarms most often went to the environment that was most similar to the original environment. For example source memory was most often confused between the two indoor house environments and the two farm environments, which is illustrated by the darker boxes in Figure 3.5. We therefore conducted an exploratory analysis to test the hypothesis that congruent sounds improved general source memory for the environment category, i.e., the object was encoded in an indoor (house) or outdoor (farm) environment. An exploratory, one-tailed, paired-samples t-test showed no significant effect of Object Condition on indoor/outdoor Source Memory (percent correct), $t(67) = 0.73$, $p = 0.77$, *Cohen's d* = 0.09, $BF_{01} = 2.92$. The contrast between this test and the significant difference in specific Environment Source Memory reported above suggests that congruent sounds improved memory for the specific environment an object was encoded in rather than a

more general memory for whether an object was encoded in an indoor or outdoor environment, and the Bayes Factor provides anecdotal evidence for this null hypothesis.

Confidence. To assess whether sound congruence affected memory confidence, we analyzed average confidence ratings in memory across Object Conditions. A one-tailed, paired-samples t-test on confidence ratings showed a significant effect of Object Condition on memory Confidence, $t(67) = 1.66$, $p = 0.050$, *Cohen's d* = 0.113, such that confidence was higher for objects paired with congruent sounds at encoding ($M = 3.32$, $SD = 0.72$) than for objects paired with control sounds at encoding ($M = 3.23$, $SD = 0.76$). The Bayes Factor provides anecdotal evidence for the null hypothesis for this effect ($BF_{01} = 0.98$; Figure 3.4b).

Location Memory. A one-tailed, paired-samples t-test on memory for the specific object in the Location Judgement (*Euclidean distance in pixels*) did not show a significant effect of Object Condition, $t(67) = 0.64$, $p = 0.26$, *Cohen's d* = 0.06; the distance between the true object location and the reported location was not different for objects encoded in the Audiovisual Congruent condition ($M = 203.12$, $SD = 187.10$) than for objects encoded in the Audiovisual Control condition ($M = 226.81$, $SD = 184.18$) (Figure 3.4c). While performance between these two conditions was not significant, this pattern of results is consistent with our hypothesis that memory for object location would be more precise for objects encoded with a congruent sound than a meaningless control sound. The Bayes Factor suggests only anecdotal evidence for the null hypothesis between these groups ($BF_{01} = 3.03$). Results from the Map Orientation Task showed overall higher performance ($M = 118.07$, $SD = 105.72$) than on Location Memory, suggesting that they were able to perform the translation from scene-view to map-view more accurately when they were looking at both the scene and map concurrently. This suggests that

performance was low for Location Memory not because participants could not perform the task properly, but because the task was more difficult and/or Location Memory was not very precise.

Encoding & Memory Performance Relationship. We used a correlation analysis to explore the relationship between successful encoding task performance and subsequent recognition memory. There was not a significant correlation between encoding task performance (M = 81.42%, SD = 9.68%) and recognition task performance (M = 39.71%, SD = 8.83%; $r(68) = 0.18$, $p = 0.1127$).

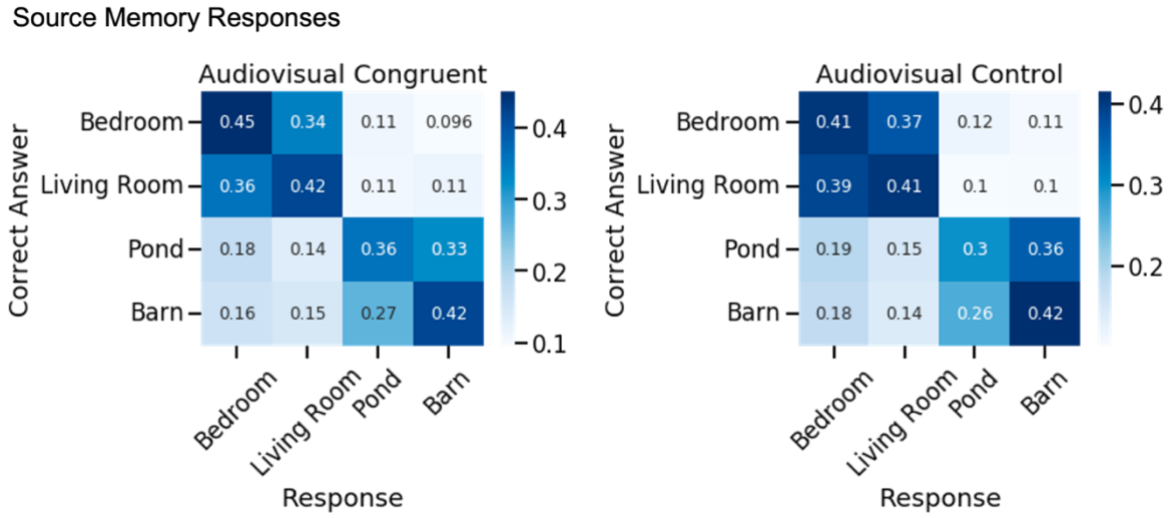


Figure 3.5. Experiment 2 Source Memory Response Matrices. These matrix heatmaps illustrate the proportions of source memory responses in each environment, with the correct environments on the y axes, and the responses on the x axes for objects encoded in the Audiovisual Congruent and Audiovisual Control conditions, with the correct responses on the diagonal, and false alarms on the off-diagonals. The darker blue diagonal for objects encoded in the Audiovisual Congruent condition (left) illustrates the more precise memory for where these objects were encoded, relative to objects encoded in the Audiovisual Control condition.

Debriefing Questionnaire. For this analysis, we assessed responses to the open-ended debriefing questionnaire prompt which asked participants to report an example of something that helped them remember the environment that objects were located in during encoding. Four participants did not provide responses to the prompt, and three participants provided two responses. All other participants provided one response, resulting in 70 responses that were used in the analysis. Of the 70 responses, six categories of information emerged, including: 1)

personal opinions of or connections to items (5.71% of responses), 2) memory for object-congruent sounds (12.86% of responses), 3) visual features of objects, such as size, color, or shape (17.14% of responses), 4) memory for items nearby the target objects (17.14% of responses), 5) congruency of an object to a scene (20% of responses), and 6) memory that an object was placed in an unusual or unexpected location (27.14% of responses; Figure 3.4d). The proportions of mentions indicate that the most salient scene features driving memory for the environment that an object was encoded in were related to how well an object fit in the environment or how surprising it was to see the item in the location that it was in. This is unsurprising given that object-scene incongruency is known to capture attention (e.g., Biederman, 1972, Henderson et al., 1999). Interestingly, nine participants (12.86%) reported that congruent sounds helped them remember where objects were located within the environment. One participant reported, "when the sound matched, like the bell in the barn, I was able to recall the location almost instantly." While sound congruency was not the most frequently reported piece of memory-supporting information, the presence of these mentions supports our hypothesis that encoding a visual object along with a congruent sound can lead to better memory not only for the object itself and the sound, but also other features of the encoding context.

Experiment 2 Discussion

The results from Experiment 2 supported our hypothesis that encoding an object along with a congruent sound facilitates source memory for features of the object's encoding context. This was demonstrated in the comparison of performance on the 4-AFC Environment Source Memory for items encoded with congruent sounds and those with meaningless control sounds. This finding is particularly striking given the task did not require attention to the sounds or the environment, and participants were unaware their memory would be tested. Self-reported

confidence ratings showed that participants were also more confident in correct responses for objects in the Audiovisual Congruent condition compared to objects in the Audiovisual Control condition. Further, we found that congruent sounds supported memory for the specific environment that the object was encoded in, rather than a more general sense of whether an object was encoded in an indoor or outdoor environment. Together, these results suggest that source memories for audiovisual objects are stronger and more precise than those for visual objects with no congruent sound information.

While we expected location judgements to be more precise for objects encoded in the Audiovisual Congruent condition, we did not find that to be the case, though performance was numerically better (smaller distance) than in the Audiovisual Control condition. We expect that this was at least in part due to very low performance on the location task, suggesting that participants made a large number of guesses. This is supported by debriefing survey responses, for which 85.29% of participants reported finding this task "very difficult," and the remaining 14.71% reported finding the task "difficult."

Overall, results were consistent with our main hypothesis, suggesting that hearing a sound when encoding a matching visual object can lead to better memory not only for the object itself, but also for its association with the environmental context in which it was encoded.

General Discussion

The goal of the present research was to investigate how multisensory object processing affects episodic memory for other object and context information present during encoding. In Experiment 1a, we found that pairing an object with a congruent sound yielded recollection benefits, without significantly improving or impairing memory for neighboring visual objects. Experiment 1b showed that increasing the task-relevance of the relation between an Audiovisual

Congruent and a Neighboring Visual object during encoding eliminated the benefit of multisensory encoding altogether suggesting that relational encoding overpowers and effectively negates the advantage typically gained from processing a multisensory object. Lastly, Experiment 2 showed that memory for the environment in which an object was encoded was better for those paired with congruent sounds than meaningless control sounds. Together, results from these studies provide evidence that audiovisual object processing impacts episodic memory by strengthening memory for that object without disrupting memory for nearby objects and by improving memory for the scene context in which the object was encoded. These findings extend previous research on multisensory memory benefits and have implications for understanding episodic memory formation in naturalistic encoding contexts in which multisensory stimuli are pervasive.

The item-based memory test used in Experiment 1a allowed us to uncover an important aspect of the multisensory memory effect, which is that it does not negatively impact memory for nearby visual items. By testing memory for items individually, rather than memory for the association between co-presented objects, we found that the memory benefit of multisensory object processing does not come at a cost to memory for individual objects co-presented alongside them. Previous work has shown better memory for items that were selectively attended during encoding, and worse memory for contextual information that was not attended (e.g., Uncapher & Rugg, 2009). While selective attention may still play a role in audiovisual memory benefits, our results suggest that if it does, it uniquely does not come at the cost to memory for other nearby objects.

Results from Experiment 2 are particularly striking given that neither the sounds nor the scene contexts were relevant to the encoding task, yet the object-congruent sounds led to better

memory for the scene context. While studies have been conducted to characterize the factors that improve associative and source memory in this way (see Yonelinas et al., 2022; Mitchell & Johnson, 2009), none, to our knowledge, have shown improved source memory when the to-be-associated information is task-irrelevant. This suggests that multisensory object processing may improve the ability to recognize objects *and* to differentiate between episodic memories for a particular object by improving memory for where that object was encoded. These results complement the results from Experiment 3 of Duarte et al. (2022), in which they found better source memory for sounds that were congruent with a simultaneously presented visual object (e.g., *ribbit* + frog compared to *ribbit* + dog). In that study, the sounds were not relevant to the encoding task. Murray et al. (2022) took the associative memory test a step further to test memory for faces paired with visually or audiovisually presented names, though the associations were still in reference to a single object (a person), rather than a multisensory object plus some additional context. Experiment 2 of the present research shows that this source memory benefit extends to information outside of the object itself to support memory for novel associations between congruent audiovisual objects and the context in which they appeared. Together, our results and those of previous studies characterize the multisensory memory effect as impacting memory for associations directly related to the constituent stimuli (i.e., the visual and auditory elements of a single object; Duarte et al., 2022), associations between faces and names (Murray et al., 2022), and incidentally encoded contextual scene information (the present study).

Proposed mechanisms underlying the multisensory memory effect are based largely on behavioral evidence for the robust memory representations created by audiovisual encoding. According to Dual Coding theory, audiovisual processing improves memory by building a memory trace with multiple paths to access that representation (Thompson & Paivio, 1994).

They posited that auditory and visual codes of a single stimulus (i.e., a spoken and written word) are stored separately in memory and function as unique encodings of a single item, providing two independent memory codes that can be relied on for a single object. However, Murray et al. (2022) found that the audiovisual benefit for associative memory for faces and names was due to multisensory integration rather than simply having multiple independent memory traces, such that the benefit was greater when the audiovisual pairs were integrated into a single percept by comparing memory for synchronously versus asynchronously encoded audiovisual stimuli. This suggests that the creation of an integrated multisensory percept provides associative memory benefits over-and-above the benefits of just having a second redundant route to accessing the memory representation, and that the constituents of the audiovisual pair bind to the other information present at encoding as well. In their study, the audiovisual and contextual information were all associated with the same object (i.e., names, faces, and voices), but our study demonstrates how their multisensory encoding framework could be extended to show that the rich encoding of audiovisual objects benefits other objects as well.

While much of the previous research on multisensory memory has focused on individual objects and their corresponding sounds, there has been some work assessing the effects of sounds in memory for dynamic video clips. Some research has found a memory benefit for dynamic movie clips to be greater than static unisensory snapshots (Buratto et al., 2009; Matthews et al., 2007; Matthews et al., 2010). Others have shown additional benefits of audiovisual dynamic rather than visual dynamic scenes (Meyerhoff et al., 2023; Meyerhoff & Huff, 2016). The study by Meyerhoff & Huff (2016) specifically found that audiovisual film clips were better remembered than unimodal visual or auditory clips, even when the auditory and visual stimuli are offset temporally, which they interpret as implying a distinct audiovisual integration process

for scenes or dynamic information that relies on learned semantic associations more than low-level factors like temporal synchrony. This contrasts the findings of Murray et al. (2022), which did find a benefit of temporal synchrony between audiovisual stimuli at encoding on memory. This opens the possibility that there are distinct or multiple mechanisms by which audiovisual processing supports memory for individual objects and for episodic memories. Within the Meyerhoff & Huff study, and Experiment 2 of our study, it is possible that multisensory integration was not the main driver of the memory benefit, but rather the presentation of a sound that reinforced the visual identity. This could be addressed using a similar paradigm to our Experiment 2 with redundant unimodal information contrasted with redundant crossmodal information. Future research in this area will provide valuable insight into the potentially multifaceted impacts of multimodal information on memory for objects and events, which will be especially important for generalizing laboratory findings to the real world.

The benefit of multisensory object encoding on object-context binding and association building found in our study also has implications for understanding the mechanisms underlying episodic memory encoding in general. There is a great deal of evidence that memory for objects and object-context associations are behaviorally distinguishable and are supported by distinct neural mechanisms. For example, evidence from neuroimaging and patient studies suggests that the hippocampus plays an integral role in binding objects to their context to form the basis of episodic memories, while other areas of the medial temporal lobe, such as the perirhinal cortex, are involved in memory for individual objects (Davachi & Wagner, 2002; Ekstrom & Ranganath, 2018; Yonelinas et al., 2010, 2019; Diana et al., 2007). Duarte et al. (2022) found the multisensory memory effect to be specific to recollection-based recognition memory, with no effect on individual item familiarity, and results from the present study and Murray et al. (2022)

provide evidence that multisensory processing at encoding facilitates the learning of novel associations. Given the role of the hippocampus in recollection, associative learning, and object-context binding, these studies suggest that the hippocampus may play a unique role in encoding multisensory objects. There is some evidence suggesting that the hippocampus is especially involved in creating novel associations across modalities (Borders et al., 2017; Kok & Turk-Browne, 2018). For example, Borders et al. (2017) had participants with hippocampal and broader MTL lesions study novel audio-visual and unisensory (visual-visual and audio-audio) stimulus pairs. Compared to healthy controls, both patient groups were significantly more impaired in their associative memory for audio-visual pairs, suggesting that the hippocampus is particularly necessary for building cross-modal associations. This lends credence to the hypothesis that the hippocampus plays a unique role in encoding audiovisual objects, though more research is necessary to understand how this region and other MTL regions contribute to audiovisual object-context binding. Studying these neural mechanisms will be important for understanding why multisensory object encoding impacts memory for context outside of the object itself, and will contribute to building ecologically valid models of episodic memory.

In summary, the present research provides novel evidence that encoding congruent audiovisual objects improves memory for both the audiovisual object and the environmental context in which it was encoded, and that the benefit does not come at a cost to memory for individual objects encoded nearby. This study extends previous work showing multisensory memory benefits to recognition memory (Duarte et al., 2022; Heikkilä et al., 2015, 2017; Lehmann & Murray, 2005; Matusz et al., 2017; Murray et al., 2022; Thelen et al., 2015; Thelen & Murray, 2013) to better our understanding of how multisensory processing impacts the formation of episodic memories. In more naturalistic settings, multiple mechanisms may work in

tandem to produce unique memory outcomes for audiovisual objects and events. Because multisensory processing has also been shown to increase stimulus detection and attentional capture, these mechanisms may impact memory during real-world experiences. Understanding the interplay between these processes will ultimately help us uncover how memories are formed in real-world situations, in which visual objects and events are object encoded in the context of other, multimodal information. Further, while other encoding strategies such as elaborative processing, increased attention, and mnemonic building have been shown to increase recollection or associative learning, they all require effortful shifts in controlled strategies. In contrast, our study demonstrates that the multisensory encoding results in memory benefits even under incidental encoding conditions. Further investigating the parameters of the multisensory memory effect and its underlying mechanisms will provide insights into memory systems and opportunities for the development of new strategies for improving memory and learning.

Chapter 3 Supplemental Materials

Supplemental Table 3.1. Stimuli for Experiment 1a, sounds also used across experiments.

Old/New	Object Size	Object	Sound Description
Old	Small	cup with straw	*Slurping from near-empty cup through straw*
		snake	*Hiss*
		toaster	*Toaster pop*
		keyboard	*Keyboard typing*
		spray bottle	*Spray mist*
		bird	*Bird chirping*
		light switch	*Click of switch turning on/off*
		coins	*Coins chiming as they hit one another*
		key	*Key jangling against one another*
		tennis racket	*Racket hitting tennis ball*
		camera	*Film camera shutter*
		harmonica	*Note played on harmonica*
		soda can	*Can being opened*
		smartphone	*Phone ringtone*
		xylophone	*Notes ascending on xylophone*
		ping pong paddle	*Paddle hitting ball*
		bowling pin	*Pins falling onto hardwood*
		bat	*Bat hissing/chirp*
		flute	*Note played on flute*
		chick	*Baby chick chirping*
		rat	*Rat squeaking*
		clock	*Ticking clock*
		matches	*Matchstick strike*
		stapler	*Stapling paper*
		cat	*Meow*
		kettle	*Water boiling and bubbling*
		blender	*Whirring*
		hairdryer	*Air blowing*
		tape	*Tape being cut from roll*
		drill	*Electric drill whirring*
		laptop	*Windows startup sound*
		hammer	*Striking wood*
		hen	*Chicken clucking*
		scissors	*Snipping*
		pencil	*Writing on paper*
		teapot	*Teapot whistle*
		cup	*Liquid slurped from mug*
		book	*Pages turning*
		wine bottle	*Cork popping out*
		saw	*Sawing wood*
		maracas	*Rattling*
		basketball	*Basketball bouncing*
		lighter	*Lighter clicking on*
		frog	*Ribbit*
		bell	*Ding*
	Large	fireplace	*Fire crackling*
		tiger	*Tiger roar*
		washer	*Clothes tumbling and low engine whirring*
		cymbal	*Cymbal crash*

		axe	*Striking wood*
		boat	*Boat horn*
		jet	*Jet engine running*
		helicopter	*Helicopter blades spinning and engine running*
		drum	*Snare drum being hit*
		printer	*Document printing*
		guitar	*Notes strummed on guitar*
		door	*Creaking open/closed*
		leopard	*Leopard hiss*
		goat	*Bleating*
		crocodile	*Growl*
		golfclub	*Striking golf ball*
		car	*Car engine turning on*
		horse	*Whinny*
		billiards	*Billiards balls crashing into each other*
		dog	*Bark*
		cow	*Moo*
		arcade game	*Video game beeps*
		anvil	*Clank of hammer hitting anvil*
		motorcycle	*Revving engine*
		bear	*Bear roar*
		sink	*Faucet running and turning off*
		train	*Train running along tracks*
		baseball bat	*Bat hitting baseball*
		skateboard	*Skateboard rolling on concrete*
		bicycle	*Bicycle bell*
		toilet brush	*Scrubbing*
		bow & arrow	*Release of arrow*
		sled	*Sliding on snow*
		chair	*Sliding along wood floor*
		filing cabinet	*Sliding along tracks into closed position*
		toilet	*Flushing*
		microwave	*Beeping upon completion*
		wolf	*Howl*
		elephant	*Elephant wail*
		penguin	*Penguin chirps*
		deer	*Deer bleating*
		pig	*Oinks*
		piano	*Notes played on piano*
		sword	*Sword unsheathing and hitting metal*
		goose	*Goose honk*
New		wrench	
		whale	
		watering can	
		walkie talkie	
		scooter	
		shark	
		salamander	
		record player	
		rhino	
		rabbit	
		fridge	
		fish	

		crab	
		frypan	
		camel	
		butterfly	
		bus	
		radio	
		air hockey table	
		briefcase	
		calculator	
		candle	
		football	
		hoe	
		lamp	
		microphone	
		notepad	
		plant	
		pliers	
		scorpion	
		screw	
		screwdriver	
		snail	
		soccer ball	
		spider	
		spoon	
		table	
		turtle	
		umbrella	
		water bottle	
		whisk	
		eraser	
		octopus	
		pen	
		toucan	
		zebra	
		bed	
		cake stand	
		closet	
		couch	
		espresso maker	
		mokapot	
		stove	
		barrel	
		bucket	
		crowbar	
		rope	
		scale	
		tv	
		bedside table	
		mirror	
		tree	
		bathtub	
		spatula	
		basket	

		box	
		duffel	
		giftbox	
		pallet	
		shelf	
		suitcase	
		trash can	
		treasure chest	
		extinguisher	
		tire	
		chess	
		cone	
		streetlight	
		hydrant	
		ladder	
		mailbox	
		street sign	
		beach chair	
		fork	
		life ring	
		outlet	
		shovel	
		surfboard	
		clipboard	
		globe	

Supplemental Table 3.2. Visual object pairs (object 1 + object 2) for Experiment 1b. All listed items were used as old items, new items were the same as those in Experiment 1a (see Table 1).

Object Relatedness	**Object 1**	**Object 2**
Related	drum	guitar
	stapler	tape
	key	coins
	chair	door
	xylophone	maracas
	filing cabinet	printer
	wolf	deer
	snake	rat
	laptop	smartphone
	hammer	anvil
	bell	clock
	goat	pig
	billiards	arcade game
	cup straw	soda can
	basketball	tennis racket
	cup	teapot
	cow	hen
	kettle	blender
	sword	bow arrow
	spray bottle	sink
	skateboard	bicycle
	fireplace	axe
Unrelated	bat	ping pong paddle

	bear	chick
	dog	harmonica
	scissors	wine bottle
	hairdryer	cat
	microwave	penguin
	tiger	toilet
	light switch	frog
	saw	book
	baseball bat	matches
	keyboard	car
	horse	camera
	leopard	jet
	goose	cymbal ride
	elephant	boat
	bird	drill
	toaster	golfclub
	motorcycle	pencil
	flute	train
	lighter	bowling pin
	crocodile	piano
	helicopter	toilet brush
	washer	sled

Supplemental Table 3.3. Visual Stimuli for Experiment 2 and their Environment Pairs.

Environments	**Object 1**
Farm (barn, pond)	anvil
	basketball
	bat
	bear
	bell
	bicycle
	bird
	boat (toy)
	bow arrow
	broom
	camera
	car
	cat
	chainsaw
	chick
	coins
	cow
	cup
	cup & straw
	deer
	drill
	frog
	goat
	goose
	guitar
	hammer
	harmonica

	hen
	horse
	key
	lighter
	matches
	motorcycle
	pig
	rat
	saw
	sheep
	shovel
	snake
	soda can
	tennis racket
	train (toy)
	wolf
	wrench
House (living room, bedroom)	arcade game
	baseball bat
	billiards
	blender
	bowling pin
	clock
	crocodile (toy)
	cymbal
	dog
	drum
	elephant (toy)
	filing cabinet
	flute
	golfclub
	hairdryer
	helicopter (toy)
	jet (toy)
	kettle
	keyboard
	knife
	laptop
	leopard (toy)
	light switch
	maracas
	microwave
	pencil
	penguin (toy)
	piano
	ping pong paddle
	printer
	scissors
	skateboard
	sled
	smartphone
	stapler
	sword

	tape
	teapot
	tiger (toy)
	toaster
	wall alarm
	washer
	wine bottle
	xylophone

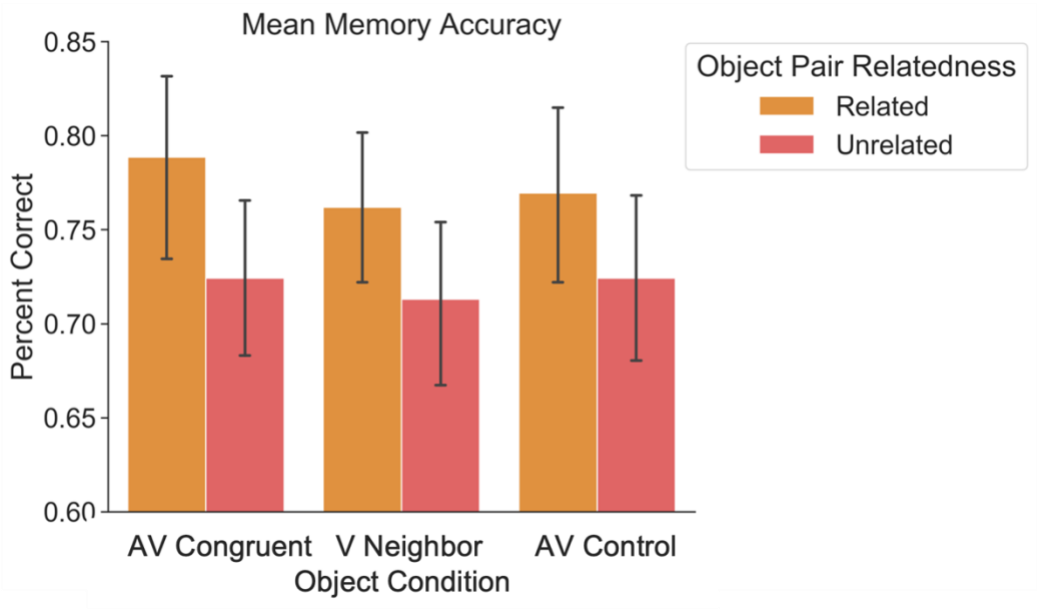
Supplemental Table 3.4. Debriefing questions for Experiments 1a, 1b, and 2

Experiment	Question	Response Options
1a, 1b, 2	Was the volume on your computer enabled throughout the entire first task?	Yes; No
	Did you adjust your volume at any time during the experiment?	Yes; No
	How would you describe the volume of the sounds during the first task?	Quiet; Loud; Just right; I did not hear sounds
	Did you use external speakers, in-ear headphones, over-ear headphones?	External speakers; In-ear headphones; Over-ear headphones; Other; I did not hear sounds
	How would you describe the environment in which you completed the study?	Quiet; Mostly quiet; Somewhat noisy; Very noisy
	Did you experience distractions during the study?	Yes, major distractions; Yes, minor distractions; No
	How much effort did you put into the experiment?	Not any; Not very much; Some effort; A lot of effort
	How difficult did you find task 1?	Not difficult; A little difficult; Very difficult
	How difficult did you find task 2?	Not difficult; A little difficult; Very difficult
	In the first task, did you ever experience a lag or gap between when the picture was shown and when the sound started?	Yes, a couple of times; Yes, often or always; No; I did not hear sounds
1a, 1b	In the second task, did you understand when you were supposed to press the "Recollect" button?	Yes, definitely; Yes, I think so; Not sure; Not at all
	Please give an example of something you recalled about an object on a trial where you pressed the "Recollect" button.	Free response
2	Give an example of something that helped you remember items better in the memory test.	Free response
	Did you find it difficult to remember which environment you saw each item?	Not difficult; A little difficult; Very difficult
	Did you find it difficult to remember where in the environment you saw each item?	Not difficult; A little difficult; Very difficult
	Were there any tasks or parts of the study that you put less effort into? Briefly explain.	Free response

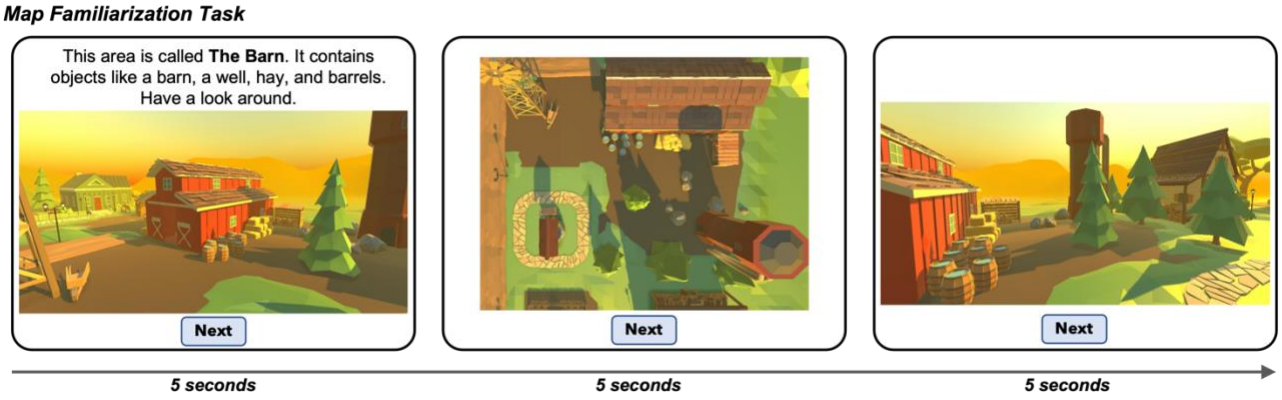
Supplemental Table 3.5. Experiment 1b additional results. Mean accuracy (% correct recognitions of old items) for Experiment 1b for items that were encoded as part of related and unrelated pairs for each Object Condition.

Object Condition	Related	Unrelated
Audiovisual Congruent	78.85(17.04)	72.41(15.01)
Visual Neighbor	76.19(15.02)	71.30(16.53)
Audiovisual Control	76.95(18.25)	72.42(16.40)

Standard deviations are shown in parentheses.



Supplemental Figure 3.2. Accuracy (% correct recognitions of old items) for Experiment 1b for items that were encoded as part of related and unrelated pairs for each Object Condition. For a list of related and unrelated pairs, see Table 1. Error bars denote standard error of the mean.



Supplemental Figure 3.2. Map Familiarization task used in Experiment 2. Participants were introduced to the environments with written descriptions, overhead maps, and images of each environment. Each image was presented for five seconds, after which participants could advance at their own pace. They were instructed to familiarize themselves with these environments because they would be used in subsequent tasks.

Supplemental Table 3.6. Demographic information for the entire sample including excluded participants for each experiment.

Experiment	Total Participants	Mean Age	Female Count	Male Count	Nonbinary Count
1a	109	19.27	77	29	3
1b	83	19.83	65	17	1
2	209	19.35	152	52	5

Supplemental Table 3.7. Race and Ethnicity information for the entire sample including excluded participants for each experiment.

Experiment	Racial Category	Hispanic/Latino	Not Hispanic/Latino	Total
1a	American Indian/Alaska Native	1	0	1
	Asian	0	61	61
	Native Hawaiian or Other Pacific Islander	1	2	3
	Black or African American	0	2	2
	White	7	16	23
	More Than One Race	2	4	6
	Not Listed	12	1	13
1b	American Indian/Alaska Native	1	0	1
	Asian	0	46	46
	Native Hawaiian or Other Pacific Islander	0	0	0
	Black or African American	0	0	0

	White	9	12	21
	More Than One Race	3	1	4
	Not Listed	10	1	11
2	American Indian/Alaska Native	0	0	0
	Asian	2	130	132
	Native Hawaiian or Other Pacific Islander	0	1	1
	Black or African American	0	6	6
	White	12	27	39
	More Than One Race	5	9	14
	Not Listed	15	2	17

Supplemental Table 3.8. Encoding Performance in each experiment. Mean performance is reported, with standard deviations in parentheses. Overall mean performance is reported, as well as performance split between trials with meaningful sounds (Audiovisual) and control white noise sounds (Control sound)

Experiment	Task	Trial Type (Condition)	Performance (% Correct)
1a	Retrocue Size Judgement Task	Overall	60.27% (5.60%)
		Audiovisual	61.40% (7.19%)
		Control sound	58.00% (8.43%)
1b	Relational Judgement Task	Overall	75.93% (18.29%)
		Audiovisual	75.47% (17.92%)
		Control sound	76.40% (18.82%)
2	Size Judgement Task	Overall	81.42% (9.68%)
		Audiovisual	81.85% (9.21%)
		Control sound	80.98% (10.17%)

Appendix A: Remember/Know Analysis

Methods.

Analysis Description. To assess whether our ROC results converge with another common method of recollection and familiarity process dissociation, we included an explicit measure of recollection-based responses for Experiments 1a and 1b (the "recollect" response option) so that we could conduct an analysis in accordance with the remember/know (recollect/familiar) procedure (Tulving, 1985). For this, we assessed recollection/remember accuracy under a

threshold assumption by subtracting incorrect "recollect" responses from correct "recollect" responses. We did this for each participant and each Object Condition, and compared recollection performance between Object Conditions using repeated measures analyses of variance (RM ANOVAs) for each experiment and post-hoc pairwise t-tests for significant F-tests. To assess familiarity/know-based recognition, we combined responses to the "definitely old," "probably old," "maybe old," as "old" responses and combined "definitely new," "probably new," "maybe new," as "new" responses. We used these values to compute the hit rates (the proportion of correct "old" responses to old items out of the total number of old items) for each participant and each Object Condition, and the false alarm rates (the proportion of incorrect "old" responses to new items out of the total number of new items) for each participant. We then calculated d-prime for each participant and each Object Condition by subtracting the z-scored false alarm rate from the z-scored hit rates, and then we performed RM ANOVAs for each experiment to compare familiarity-based recognition (d-prime) between Object Conditions.

Results. Experiment 1a: A RM ANOVA showed a significant effect of Object Condition on recollection (remember) responses, $F(2, 98) = 6.21$, $p = 0.005$, $\eta^2 = 0.015$, with more recollect responses for items in the Audiovisual Congruent condition than in the Audiovisual Control condition $t(49) = 3.47$, $p = 0.003$, and in the Neighboring Visual than in the Audiovisual Control Condition $t(49) = 2.86$, $p = 0.02$, but no significant difference between items in the Audiovisual Congruent and Neighboring Visual Conditions, $t(49) = 1.02$, $p = 0.94$. A RM ANOVA showed no significant effect of Object Condition on familiarity (know/d-prime) $F(2, 98) = 0.70$, $p = 0.49$.

Experiment 1b: A RM ANOVAs showed no significant effects of Object Condition on either recollection (remember) responses, $F(2, 98) = 1.70$, $p = 0.19$, or familiarity (know/d-prime), $F(2, 98) = 0.52$, $p = 0.60$.

These patterns of results converge with the ROC analyses reported in the article, with the exception of the significant difference in this remember/know analysis for Experiment 1a between the objects in the Visual Neighbor and Audiovisual Control Conditions. This provides evidence that there is a subjective difference between memories for objects in these conditions, even though this is not reflected in the recollection metric (*y-intercept*) in the ROC analysis for Experiment 1a.

References

Audacity Team (2021). Audacity(R): Free Audio Editor and Recorder [Computer application]. Version 3.0.0 retrieved March 20th, 2022 from https://audacityteam.org/.

Biederman, I. (1972). Perceiving real-world scenes. Science, 177(4043), 77–80. https://doi.org/10.1126/science.177.4043.77

Borders, A. A., Aly, M., Parks, C. M., & Yonelinas, A. P. (2017). The hippocampus is particularly important for building associations across stimulus domains. *Neuropsychologia*, *99*, 335–342. https://doi.org/10.1016/j.neuropsychologia.2017.03.032

Brady, T. F., Robinson, M. M., Williams, J. R., & Wixted, J. (2021). *Measuring memory is harder than you think: A crisis of measurement in memory research* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/qd75k

Buratto, L. G., Matthews, W. J., & Lamberts, K. (2009). Short article: When are moving images remembered better? Study–test congruence and the dynamic superiority effect. *Quarterly Journal of Experimental Psychology*, *62*(10), 1896–1903. https://doi.org/10.1080/17470210902883263

Davachi, L., & Wagner, A. D. (2002). Hippocampal Contributions to Episodic Encoding: Insights From Relational and Item-Based Learning. *Journal of Neurophysiology*, *88*(2), 982–990. https://doi.org/10.1152/jn.2002.88.2.982

Diana, R. A., Yonelinas, A. P., & Ranganath, C. (2007). Imaging recollection and familiarity in the medial temporal lobe: A three-component model. *Trends in Cognitive Sciences*, *11*(9), 379–386. https://doi.org/10.1016/j.tics.2007.08.001

Dickerson, B. C., & Eichenbaum, H. (2010). The Episodic Memory System: Neurocircuitry and Disorders. *Neuropsychopharmacology*, *35*(1), 86–104. https://doi.org/10.1038/npp.2009.126

Duarte, S. E., Ghetti, S., & Geng, J. J. (2022). Object memory is multisensory: Task-irrelevant sounds improve recollection. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-022-02182-1

Eichenbaum, H., Yonelinas, A. P., & Ranganath, C. (2007). The Medial Temporal Lobe and Recognition Memory. *Annual Review of Neuroscience*, *30*(1), 123–152. https://doi.org/10.1146/annurev.neuro.30.051606.094328

Ekstrom, A. D., & Ranganath, C. (2018). Space, time, and episodic memory: The hippocampus is all over the cognitive map. *Hippocampus*, *28*(9), 680–687. https://doi.org/10.1002/hipo.22750

Fisher, R. P., & Craik, F. I. M. (1980). The effects of elaboration on recognition memory. *Memory & Cognition*, *8*(5), 400–404. https://doi.org/10.3758/BF03211136

Heikkilä, J., Alho, K., Hyvönen, H., & Tiippana, K. (2015). Audiovisual Semantic Congruency During Encoding Enhances Memory Performance. *Experimental Psychology*, *62*(2), 123–130. https://doi.org/10.1027/1618-3169/a000279

Heikkilä, J., Alho, K., & Tiippana, K. (2017). Semantically Congruent Visual Stimuli Can Improve Auditory Memory. *Multisensory Research*, 13.

Henderson, J. M., Weeks, P. A., Jr., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. Journal of Experimental Psychology: Human Perception and Performance, 25(1), 210–228. https://doi.org/10.1037/0096-1523.25.1.210

K Kinsbourne, M., & Warrington, E. K. (1962). The effect of an aftercoming random pattern on the perception of brief visual stimuli. The Quarterly Journal of Experimental Psychology, 14(4), 223–234. https://doi.org/10.1080/17470216208416540

Kok, P., & Turk-Browne, N. B. (2018). Associative Prediction of Visual Shape in the Hippocampus. *The Journal of Neuroscience*, *38*(31), 6888–6899. https://doi.org/10.1523/JNEUROSCI.0163-18.2018

Lehmann, S., & Murray, M. M. (2005). The role of multisensory memories in unisensory object discrimination. *Cognitive Brain Research*, *24*(2), 326–334. https://doi.org/10.1016/j.cogbrainres.2005.02.005

Macaluso, E., Noppeney, U., Talsma, D., Vercillo, T., Hartcher-O'Brien, J., & Adam, R. (2016). The Curious Incident of Attention in Multisensory Integration: Bottom-up vs. Top-down. *Multisensory Research*, *29*(6–7), 557–583. https://doi.org/10.1163/22134808-00002528

Matthews, W. J., Benjamin, C., & Osborne, C. (2007). Memory for moving and static images. *Psychonomic Bulletin & Review*, *14*(5), 989–993. https://doi.org/10.3758/BF03194133

Matthews, W. J., & Buratto, L. G. (n.d.). *Exploring the memory advantage for moving scenes*.

Matusz, P. J., Wallace, M. T., & Murray, M. M. (2017). A multisensory perspective on object memory. *Neuropsychologia*, *105*, 243–252. https://doi.org/10.1016/j.neuropsychologia.2017.04.008

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. https://doi.org/10.3758/s13421-015-0575-6

Meyerhoff, H. S., Jaggy, O., Papenmeier, F., & Huff, M. (2023). Long-term memory representations for audio-visual scenes. *Memory & Cognition*, *51*(2), 349–370. https://doi.org/10.3758/s13421-022-01355-6

Mickes, L., Hwe, V., Wais, P. E., & Wixted, J. T. (2011). Strong memories are hard to scale. *Journal of Experimental Psychology: General*, *140*(2), 239–257. https://doi.org/10.1037/a0023007

Mitchell, K. J., & Johnson, M. K. (2009). Source monitoring 15 years later: What have we learned from fMRI about the neural mechanisms of source memory? *Psychological Bulletin*, *135*(4), 638–677. https://doi.org/10.1037/a0015849

Moran, Z. D., Bachman, P., Pham, P., Hah Cho, S., Cannon, T. D., & Shams, L. (2013). Multisensory Encoding Improves Auditory Recognition. *Multisensory Research*, *26*(6), 581–592. https://doi.org/10.1163/22134808-00002436

Murray, C. A., Tarlow, M., Rissman, J., & Shams, L. (2022). Multisensory encoding of names via name tags facilitates remembering. *Applied Cognitive Psychology*, *36*(6), 1277–1291. https://doi.org/10.1002/acp.4012

Schneider, T. R., Engel, A. K., & Debener, S. (2008). Multisensory Identification of Natural Objects in a Two-Way Crossmodal Priming Paradigm. *Experimental Psychology*, *55*(2), 121–132. https://doi.org/10.1027/1618-3169.55.2.121

Shams, L., & Seitz, A. R. (2008). Benefits of multisensory learning. *Trends in Cognitive Sciences*, *12*(11), 411–417. https://doi.org/10.1016/j.tics.2008.07.006

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149. https://doi.org/10.3758/BF03207704

Stein, B. E., Stanford, T. R., & Rowland, B. A. (2020). Multisensory Integration and the Society for Neuroscience: Then and Now. *The Journal of Neuroscience*, *40*(1), 3–11. https://doi.org/10.1523/JNEUROSCI.0737-19.2019

Thelen, A., & Murray, M. M. (2013). The Efficacy of Single-Trial Multisensory Memories. *Multisensory Research*, *26*(5), 483–502. https://doi.org/10.1163/22134808-00002426

Thelen, A., Talsma, D., & Murray, M. M. (2015). Single-trial multisensory memories affect later auditory and visual object discrimination. *Cognition*, *138*, 148–160. https://doi.org/10.1016/j.cognition.2015.02.003

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), 380–398. https://doi.org/10.1037/1196-1961.48.3.380

Uncapher, M. R., & Rugg, M. D. (2009). Selecting for Memory? The Influence of Selective Attention on the Mnemonic Binding of Contextual Information. *Journal of Neuroscience*, *29*(25), 8270–8279. https://doi.org/10.1523/JNEUROSCI.1043-09.2009

Yonelinas, A. P. (2002). The Nature of Recollection and Familiarity: A Review of 30 Years of Research. *Journal of Memory and Language*, *46*(3), 441–517. https://doi.org/10.1006/jmla.2002.2864

Yonelinas, A. P., Aly, M., Wang, W.-C., & Koen, J. D. (2010). Recollection and familiarity: Examining controversial assumptions and new directions. *Hippocampus*, *20*(11), 1178–1194. https://doi.org/10.1002/hipo.20864

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. Journal of Experimental Psychology: Learning, Memory, and Cognition, 20(6), 1341–1354. https://doi.org/10.1037/0278-7393.20.6.1341

Yonelinas, A. P., & Jacoby, L. L. (1994). Dissociations of processes in recognition memory: Effects of interference and of response speed. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *48*(4), 516–535. https://doi.org/10.1037/1196-1961.48.4.516

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin*, *133*(5), 800–832. https://doi.org/10.1037/0033-2909.133.5.800

Yonelinas, A.P., Ramey, M.M., & Riddell, C. (2022). Recognition Memory: The Role of Recollection and Familiarity. In Kahana, M. & Wagner, A.D. (Eds.), Handbook of Human Memory: Foundations and Applications. Oxford University Press.

Yonelinas, A. P., Ranganath, C., Ekstrom, A. D., & Wiltgen, B. J. (2019). A contextual binding theory of episodic memory: Systems consolidation reconsidered. *Nature Reviews Neuroscience*, *20*(6), 364–375. https://doi.org/10.1038/s41583-019-0150-4

Vallat, R. (2018). Pingouin: statistics in Python. *Journal of Open Source Software*, 3(31), 1026, https://doi.org/10.21105/joss.01026

Van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., & Yu, T. (2014). scikit-image: image processing in Python. *PeerJ*, *2*, e453.

Chapter 4: Realistic stimulus properties impact audiovisual spatial ventriloquism in virtual reality

The following chapter consists of a manuscript prepared for submission for publication at *Attention, Perception, & Psychophysics*

Abstract

The ventriloquist effect occurs when the perceived location of a sound is mislocalized towards a coincidental visual stimulus. This most famously occurs when a puppet is perceived as speaking instead of its puppeteer. This illusion illustrates how our daily percepts require the active integration of visual and auditory signals. Most research on the ventriloquist effect has used simple stimuli, making it difficult to know how well laboratory findings generalize to real-world crossmodal stimuli. In two experiments, we investigate the influence of stimulus realism and dynamic motion on audiovisual integration within a ventriloquist paradigm in virtual reality (VR). In Experiment 1, we assessed spatial ventriloquism with three types of stimuli: simple static spheres paired pure tones, static realistic objects paired with semantically congruent sounds, or animated realistic objects paired with semantically congruent sounds. We found that realistic animated objects paired with congruent sounds led to an increase in ventriloquism at small spatial disparities. Bayesian causal inference modeling suggested that this is because dynamic motion coupled with semantically congruent information increases the prior of common cause that the stimuli originated from the same source when they are near each other. In Experiment 2, we found that the animated objects no longer led to greater ventriloquism when paired with a meaningless tone, indicating that it is not the dynamic motion alone, but the congruency of the dynamic object with a semantically congruent sound that enhanced the

ventriloquist effect. Together, our results suggest that realistic, animated stimulus pairs can increase audiovisual spatial ventriloquism.

Introduction

Ventriloquism is the illusion where the sound of the puppeteer's voice is mis-localized and appears to come from the moving mouth of a nearby puppet. This illusion is a classic example of multisensory integration in which a sound is perceived as originating from the location of a nearby visual stimulus. The mis-localization occurs because vision typically has greater spatial precision than auditory information and therefore the brain relies on the former to locate the latter (Knill & Pouget, 2004). Lab studies using simple stimuli, such as single-point LED lights and simple tones placed in real space, have shown that the magnitude of the ventriloquist effect depends on how closely the audiovisual stimuli correspond in space and time, and the observer's prior expectation that the two share a common source (Körding et al., 2007; Rohe & Noppeney, 2015; Slutsky & Recanzone, 2001; Van Wanrooji et al., 2010). There is also evidence, although it remains controversial, that stimulus realism affects the strength of the effect (Bruns, 2019; Chen & Spence, 2017; Noppeney, 2021). Realistic multimodal stimuli are further defined by semantic correspondences (e.g., a common object such as a dog or kettle), and dynamic spatiotemporal correspondences (e.g., a tone at the same time that a stimulus in motion hits the ground), but the individual contributions of these two characteristics to ventriloquism for object stimuli have not yet been fully investigated within a single study. In this study, we combine both types of realism in a virtual reality (VR) environment to better understand how the strength of spatial ventriloquism is impacted by semantic and dynamic spatiotemporal correspondences between audiovisual stimuli.

Studies of multisensory integration using simplistic audiovisual stimuli have uncovered much of what we know about the ventriloquist effect. For example, Wallace et al. (2004) employed a ventriloquist task using noise bursts and light-emitting diode (LED) stimuli, and found that the ventriloquist effect was most pronounced when auditory and visual stimuli were spatially close and temporally synchronous, and when they were perceived as "unified." This demonstrated that spatial and temporal proximity are critical for enhancing multisensory integration. Typical stimuli used for studying ventriloquism include flashes of light, Gaussian clouds, and noise bursts or tones. These studies generally find that the ventriloquist effect is stronger when the auditory stimulus is more proximal to the visual stimulus, when the visual stimulus is highly reliable, and when the two stimuli are temporally coincident (see Bruns, 2019 for a review). This is thought to occur because observers tend to bind auditory and visual signals that correspond in space and time, and when the visual signal is more reliable, it is weighted more heavily in the integrated percept (Alais & Burr, 2004; Rohe & Noppeney, 2015; Wozny et al., 2010, Rohe & Noppeney, 2016). In addition to the spatial proximity and temporal coincidence of the two signals, the strength of the ventriloquist effect is also affected by the observer's prior expectation that the two stimuli should originate from a single source or different sources (Van Wanrooij et al., 2010).

The factors influencing the strength of the ventriloquist effect are formalized by Bayesian causal inference models. These models describe the ventriloquist effect as resulting from the combination of perceptual spatial estimates obtained under the assumption of a single and two separate causes, weighted by posterior beliefs that the stimuli come from common or independent sources (the prior of common cause; Körding et al., 2007). This causal inference relies on the spatial and temporal disparity between the sensory signals, though it can also be

influenced by structural characteristics, like semantic and synesthetic correspondences (Delong & Noppeney, 2021; Parise & Spence, 2008; Parise & Spence, 2009). We hypothesize that semantic and dynamic spatiotemporal realism will enhance spatial ventriloquism specifically by increasing the prior expectation that two stimuli originate from a common source because the event will have greater correspondence to natural experience (e.g., a dog with a moving mouth and the sound of a bark).

Evidence for stronger audiovisual integration with realistic stimuli comes from an early study by Jackson (1953) in which observers estimated the location of a bell sound with lights arrayed across 67.5º, or a whistle sound with a kettle releasing a cloud of steam across 90º. The objects were all real objects arrayed in a room. The results showed stronger perception of the whistle coming from the kettle than the bell from the light at all distances measured. Even at the nearest disparity, participants in the bell and light condition were almost equally likely to report the bell at its true location as at the location of the light. However, in the kettle condition, participants almost always reported the whistle as coming from the kettle even at a 30º disparity; they were equally likely to say that the sound came from the kettle as the true location even at a 90º disparity. Whether or not this finding was due solely to the semantic correspondence between the kettle and whistle is not fully clear given that these stimuli also had greater spatiotemporal dynamic correspondence than the light flashes and bell sound. Although the two stimulus conditions were not perfectly equated, the experiment does provide convincing evidence that realistic stimuli mimicking real-world events produce stronger audiovisual integration than arbitrary stimuli.

Overall, while an increase in ventriloquism with realistic stimuli is consistent with findings that semantic and perceptual realism enhance crossmodal interactions in multiple

domains such as object identification (Chen & Spence, 2010; Williams et al., 2022), speech perception (Kanaya & Yokosawa, 2011; Warren et al., 1981), and memory (Matusz et al., 2017), there is some contradictory evidence. For example, a recent study in VR found no differences in the ventriloquist effect between the image of a falling handball coupled with a realistic sound of that type of ball bouncing, and a Gaussian blur visual stimulus falling with a coincident noise burst sound (Huisman et al., 2022). This study suggested that stronger ventriloquism in some studies might be due to audiovisual synchronization and not realism. Similarly, the effect of realism on the ventriloquist aftereffect, in which audiovisual spatially disparate signals induce a recalibration (i.e. shift) of the perceived location of a subsequent auditory stimulus, has been mixed (Radeau & Bertelson, 1974, 1977, 1978; Recanzone, 1998). The inconsistent findings across these studies underscore the need for additional research investigating the precise role of stimulus realism on the ventriloquist effect.

In two experiments, we explored the influence of naturalistic stimulus realism on audiovisual integration within a ventriloquist paradigm presented in virtual reality (VR). In Experiment 1, we assessed spatial ventriloquism with three types of stimuli: non-realistic (sphere + tone), static-realistic (a static scissor + semantically congruent snipping sound), or animated-realistic (an animated scissor + semantically and spatiotemporally congruent snipping sound). The goal of Experiment 1 was to test whether ventriloquism is stronger for semantically related audiovisual than meaningless signals and whether this depends on spatiotemporal dynamic correspondence. In Experiment 2, we then tested whether the increased ventriloquism for the animated realistic stimuli was a result of spatiotemporal dynamic correspondence between these stimuli by including a condition in which meaningless sounds were paired with animated visual stimuli and aligned with the spatiotemporal dynamic structure of the animations. This

experiment served as a control for the effect of audiovisual dynamic spatiotemporal coincidence during visual animation on the strength of the ventriloquist effect. In both experiments, we fit the Bayesian causal inference model to our data to test whether the effect of semantic and/or animated realism on multisensory integration can be attributed to differences in the prior of common cause and/or the reliability of each stimulus event. This methodological approach enables us to gain a clearer understanding of the factors driving audiovisual integration with more naturalistic stimuli that we regularly encounter.

Experiment 1

Method

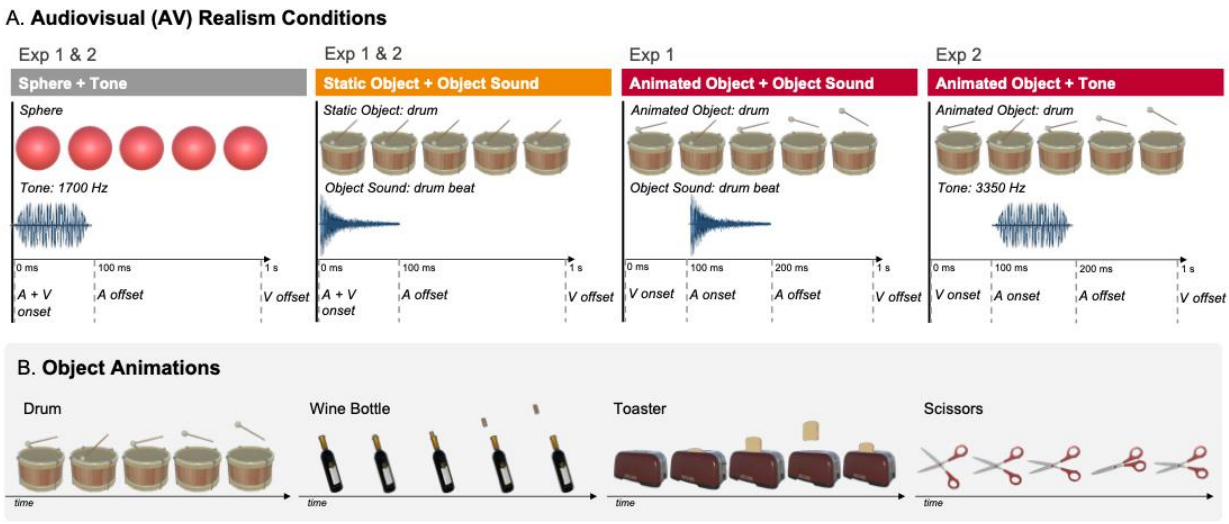


Figure 4.1. Stimuli used in Experiments 1 & 2. **A.** An example of audiovisual stimuli used in each AV Realism condition. The same Sphere + Tone condition was used in Experiments 1 & 2, and always included the static red sphere. The Static Object + Object Sound, Animated Object + Object Sound, and Animated Object + Tone conditions were comprised of the same four visual (V) stimuli, either static or animated, and the example illustrates the drum stimulus and corresponding sounds (A). **B.** This figure illustrates the animations of the four visual stimuli when used as animated objects.

Participants. Twenty-six students (20 identified as female, six identified as male, μ_{age} = 20.04 years) from the University of California, Davis, participated in exchange for partial course credit. To minimize the risk of cybersickness associated with head-mounted VR display use, participants were screened for symptom susceptibility prior to the study using the Visually Induced Motion Sickness Susceptibility Questionnaire (VIMSSQ, see Supplemental Materials;

Keshavartz et al., 2019). Three participants were excluded due to poor unisensory sound localization accuracy (see Data Analysis), one participant was excluded because they did not complete the experiment, and two participants were excluded because they did not follow instructions and reported selecting the location of the visual object instead of the sound during the audiovisual ventriloquist task.

All materials and the Unity-based experiment framework are freely available on GitHub (https://osf.io/tnwb2/; https://github.com/sheaduarte/Multisensory-Causal-Inference-VR).

Materials.

Apparatus. Stimuli were presented in an HTC Vive Eye Pro head-mounted display, and the experiment was programmed in Unity and run through Steam VR (HTC Corporation; Unity Editor Version 2019.3.8, Unity Technologies, 2019; Steam VR Version 1.6.10, Valve Corporation, 2019). To produce virtually spatialized sound signals, we used Steam Audio, which renders sounds binaurally using a generalized head-related transfer function (HRTF) to model the position of the sound relative to the participant. Although the general HRTF may not precisely suit all listeners, we confirmed that participants could reliably localize our sound stimuli before the main experiment (see Data Analysis).

Participants completed the study sitting in a chair within a 9 m by 9 m virtual room within a sound-controlled data collection room. The height of the chair was adjusted to maintain a headset height between 105 and 125 cm above the floor. Behavioral responses were recorded using a Vive controller.

Stimuli. Visual: Visual stimuli were composed of five 3-dimensional (3D) models, including a selection of realistic objects (drum, toaster, wine bottle, and scissors) obtained from the Unity 3D Asset Store, and a simple red sphere created in the Unity Editor. All visual stimuli

were edited to approximately 50 cm across in virtual size (5.71° visual angle at the presented distance of 5 m).

Auditory: Each visual stimulus had a corresponding sound obtained from FindSounds.com. The red sphere had a corresponding 100 ms, 1700 Hz tone. The sounds for the realistic objects were between 100 and 150 ms in duration and corresponded semantically with the visual objects (i.e., a drumstick hitting a snare drum, a toaster popping, a wine cork popping, and scissors snipping). The onsets of the auditory and visual stimuli were temporally synchronized for all stimuli, as further described in the Experimental Design section. The peak amplitudes of all the sounds were normalized using Audacity software (Audacity Team, 2021).

Experimental Design. Experiment 1 was designed to assess the effects of audiovisual realism on spatial ventriloquism. To this end, we employed an audiovisual ventriloquist paradigm in which we manipulated the spatial disparity between auditory and visual stimuli (AV Disparity) and the realism of the audiovisual stimulus pairs (AV Realism). Participants reported the perceived location of the sound with a virtual laser pointer emanating from a Vive controller.

AV Disparity was manipulated by presenting audiovisual stimulus pairs presented either at the same or different locations. On each trial, auditory and visual stimulus locations were sampled independently from five possible locations (angles -30º, -15º, 0º, 15º, and 30º) at a distance of 5 m along the azimuth axis, which refers to the horizontal angle of an object from the observer's viewpoint. Thus, stimuli could be presented at the same location (0° AV Disparity), or different locations (15º, 30º, 45º, 60º AV Disparity). Multiple location combinations could result in the same AV Disparity. For example, an AV Disparity of 15°, calculated as the absolute value of the visual stimulus location minus the auditory stimulus location, would result from auditory and visual stimuli presented, respectively, at -30° and -15° and at 0° and -15°. For every AV

Disparity, we used an equal number of all possible stimulus location combinations based on the five possible locations. There were two possible combinations for the 60° AV Disparity, four for the 45° AV Disparity, six for the 30° AV Disparity, and five for the 0° AV Disparity.

The AV Realism was also manipulated. The stimuli on each trial were composed of a red sphere with a 1700 hz tone (Sphere + Tone), a static realistic object with its natural sound (Static Object + Object Sound), or an animated realistic object with its natural sound (Animated Object + Object Sound). The animations were created using Unity and consisted of simple movements that were coordinated with their corresponding sound (e.g., toaster popping up, drumstick hitting drum; Figure 4.1). The drum and toaster animations were 450 ms in duration, and the scissors and wine bottle animations were 1000 ms in duration, though all objects were presented for the same amount of time (1000 ms). The animations were designed to be synchronized with their sounds (e.g., a drum beat coincident with the moment of impact between the drumstick and the drum).

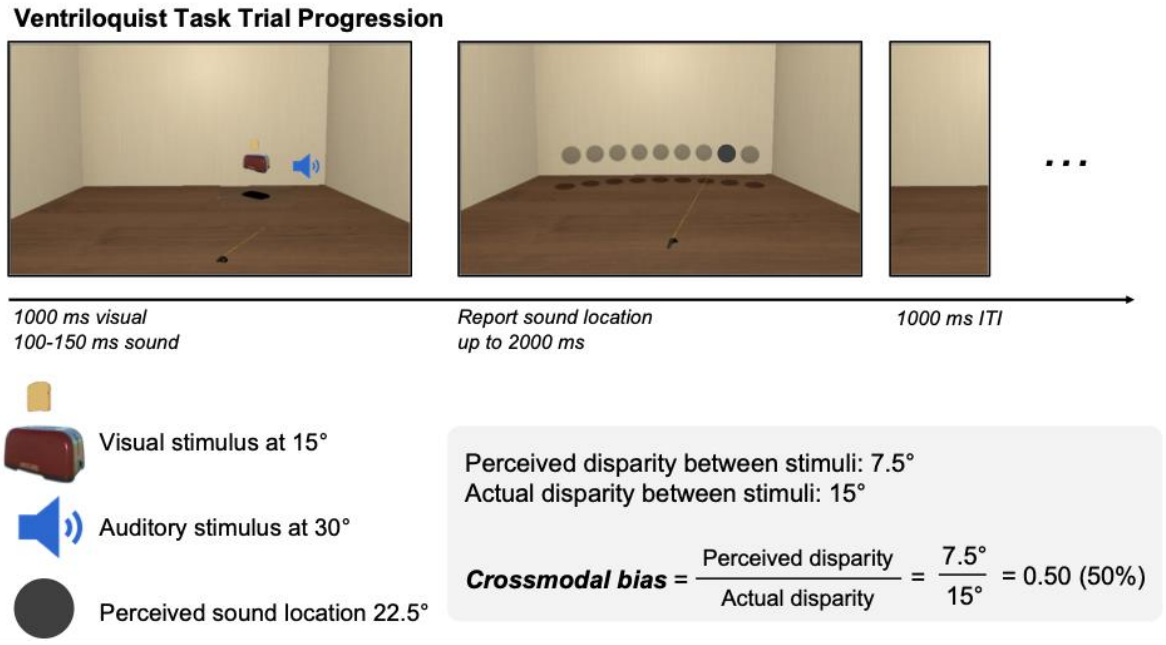


Figure 4.2. **B.** The trial progression of the audiovisual ventriloquist task, in which audiovisual stimuli were presented, followed by response options for the auditory report, presented until response (up to 2000 ms), followed

by a 1000 ms ITI. In this example, the visual stimulus (toaster) was presented at 15° visual angle, and the sound was presented at 30° visual angle, and the participant reports the sound location as 22.5° of visual angle. Because this is halfway between the visual and auditory stimuli, this demonstrates a crossmodal bias of 50%.

Procedure. On each trial, synchronous auditory and visual stimuli were presented. The visual stimulus was presented for 1000 ms, and the sound was played once through in its entirety (100-150 ms). On Sphere + Tone trials or Static Object + Object Sound trials, the sound was synchronized with the onset of the visual stimulus. On Animated Object + Object Sound trials, the sound was synchronized with the relevant point within the animation (e.g., a drum beat at the moment of impact of the drumstick on the drum) (Figure 4.1).

After the stimuli were presented, nine response options were displayed as gray spheres, located along the azimuth axis, spanning 30 degrees to both the left and right of center, at 7.5° intervals as shown in Figure 4.2. Participants were instructed to report the perceived location of the sound stimulus by pointing a laser that extended from the Vive controller to the stimulus location, and pulling the trigger to select. Though there were only 5 possible stimulus locations for a sound, nine response options were used to account for stimuli perceived to originate at an intermediate point between possible stimulus locations. Participants had a maximum of 2000 ms to respond, and a 1000 ms inter-trial interval (ITI) followed each response. Participants were informed that the sounds and visual objects could originate from the same or different locations but were not informed about the frequency of each, nor that the sound could only originate from five of the nine response options.

Participants completed 500 trials in total. There were 30 trials per condition x 3 (AV Realism: Sphere + Tone, Static Object + Object Sound, Animated Object + Object Sound) x 5 (AV Disparity: 0°, 15º, 30º, 45º, 60º) for a total of 450 trials. An additional 50 unisensory visual trials were added as attention checks. On these trials, each of the five visual stimuli was randomly sampled and presented with no sound at each of the five possible locations.

Participants were instructed to report the location of the visual stimulus. The trial order was randomized for each participant. Ten practice trials were completed prior to the main task, and these were repeated as needed for participants to understand the task. Practice trials and unisensory visual trials were not included in the analyses.

Prior to the main experiment, to ensure participants could reliably localize the visual and auditory stimuli, participants completed 25 unisensory visual localization trials followed by 50 unisensory auditory localization trials. The required response was the same as in the main experiment. Each of the five visual stimuli were presented at each of the five locations once and each of the five sounds were presented at each location twice. In total, the study took approximately 45 minutes to complete.

Data Analysis.

Unisensory localization performance. We assessed performance on the unisensory localization trials to ensure that participants could reliably localize both auditory and visual stimuli within our VR experimental setup. The low accuracy exclusion criterion was set to chance performance, 20%, on the auditory localization task with five stimulus locations. Overall accuracy (% correct) on the visual localization task was high (realistic visual objects mean = 99.5% correct, SD = 21.24%; visual sphere mean = 99.0% correct, SD = 21.32%). A paired-samples t-tests showed no significant difference in visual object localization between the realistic objects and the sphere tone. As expected, accuracy on the auditory localization task was overall lower (object sounds mean = 38.12% correct, SD = 48.60%; 1700 Hz tone mean = 31.50% correct, SD = 46.57%). A paired-samples t-tests showed no significant difference in sound localization between object sounds and the 1700 Hz tone.

In addition, we calculated the Pearson correlation between participant responses and the true signal source (e.g., Rohe & Noppeney, 2015). Consistent with reliable localization, the correlation was positive for all conditions: unisensory object sounds (across participants mean +/– SEM: 0.91 +/– 0.01), the 1700 Hz tone sound (across participants mean +/– SEM: 0.88 +/– 0.03), the realistic visual objects (across participants mean +/– SEM: 0.996 +/– 0.004), and for the sphere stimulus (across participants mean +/– SEM: 1 +/– 0.0003; see Supplemental Materials). These results demonstrate that participants could reliably locate unisensory signals across stimuli within our VR paradigm.

Mixed effects modeling of crossmodal bias. To address the primary question of whether the strength of spatial ventriloquism is affected by AV Realism and AV Disparity, we conducted a generalized linear mixed effects model (GLMM) analysis using the lme4 package in R (Bates et al., 2015). The outcome measure of interest was crossmodal bias, defined as the distance (degrees) between the reported sound location and the true sound location, divided by the disparity between the stimuli. Using stimulus disparity as the denominator normalizes the crossmodal bias to each distance, allowing us to assess the proportion of crossmodal bias at each AV Disparity as our outcome variable. Crossmodal bias was calculated for each level of AV Disparity (15°, 30°, 45°, and 60°), except for 0° because the denominator would be zero (see Figure 4.2). In addition to the crossmodal bias, there is likely also a center response bias due to the lack of response options beyond the most extreme stimulus positions (+/- 30°). However, this would be equally true of all three levels of AV Realism. Because our comparisons of primary interest were between levels AV Realism, we did not correct for center bias or any overall biases in response due to the position of stimuli across the array.

Within the model, we treated AV Realism as a categorical factor with 3 levels (Sphere + Tone, Static Object + Object Sound, Animated Object + Object Sound), and AV Disparity as continuous. The model included fixed effects of AV Disparity and AV Realism, the interaction between them, and a random effect for participants. AV Realism was dummy-coded with the Sphere + Tone condition as the reference level as it represents the baseline level of minimal realism. With this coding scheme, we can assess the effects of different levels of AV Realism on crossmodal bias relative to the Sphere + Tone reference condition. A positive beta coefficient for the Static Object + Object Sound or Animated Object + Object Sound conditions would indicate that the condition elicits more of a crossmodal bias compared to the Sphere + Tone reference, whereas a negative coefficient denotes less of a crossmodal bias for that condition.

The beta coefficient associated with AV Disparity reflects how much crossmodal bias changes with each degree of disparity in the reference condition. Specifically, a negative beta coefficient indicates that crossmodal bias decreases with increases in AV disparity. A positive beta coefficient indicates that crossmodal bias increases with each degree of AV disparity. Interactions between the Animated Object + Object Sound or Static Object + Object Sound conditions and AV disparity reflect how much more, or less, that condition changes with AV disparity compared to the reference Sphere + Tone condition. A condition with a more negative coefficient can be interpreted as having greater sensitivity to multisensory integration than the baseline condition because the likelihood of mistaking two separate AV sources as being from a single source decreases faster as the two stimuli are moved farther apart.

First, to assess whether there was an overall significant effect of AV Realism on crossmodal bias, we performed a likelihood ratio test between our full model with both predictors compared to a reduced model with only AV Disparity as a predictor. This was

conducted to understand whether adding AV Realism as a predictor led to significantly improved model fit. The likelihood ratio test indicated that the full model including AV Realism as a predictor provided a better fit for the data than a model without it (reduced model), χ^2 (8) = 50.84, $p < 0.0001$. This confirms that AV Realism affects the strength of the crossmodal bias.

$$crossmodal\ bias \sim 1 + AV\ Disparity * AV\ Realism + (1 \mid participant) \quad \text{(full model)}$$

$$crossmodal\ bias \sim 1 + AV\ Disparity + (1 \mid participant) \quad \text{(reduced model)}$$

In addition to the GLMM analysis, we conducted t-tests between AV Realism conditions at two AV Disparity levels of interest: 15° and 60°. While the GLMM was not specific to the AV Disparity levels in our study, these t-tests allow us to specifically test the differences in crossmodal bias between AV Realism conditions at the smallest and largest stimulus disparities within our experimental design. We ran Bonferroni-corrected pairwise t-tests in python using the package Pingouin (Vallat, 2018).

Bayesian Causal Inference model. To investigate how stimulus realism affects multisensory integration, we implemented a Bayesian causal inference model adapted from Körding et al. (2007). This model assesses the probability of the auditory and visual stimuli having common vs. independent causal structures by combining sensory evidence from auditory and visual stimuli with prior beliefs about their causal relationship.

The probability of a common cause ($p(C = 1)$) is determined by sampling from a binomial prior distribution of the common source prior ($P(C = 1) = p_{\text{common}}$). For a common source, the "true" location S_{AV} is drawn from the spatial prior distribution $N(\mu_P, \sigma_P^2)$, which has a mean of 0° to reflect an assumed central bias. For two independent sources, the true auditory (S_A)

and visual (S_V) locations are drawn independently from this spatial prior distribution. Auditory and visual signals are affected by sensory noise, providing imperfect information about their true underlying locations. To reflect this sensory noise, the auditory (x_A) and visual (x_V) signals are therefore drawn from normal distributions centered around the true stimulus locations ($N(\mu_A, \sigma_A^2)$ and $N(\mu_V, \sigma_V^2)$, respectively). The model therefore includes free parameters for the prior of common cause p_{common}, the variance of the spatial prior distribution (σ_P^2), and the variance of the auditory (σ_A^2) and visual (σ_V^2) distributions, representing their respective spatial reliabilities (i.e. inverse of variance).

The probability of the underlying causal structure is inferred by combining the common-source prior with the sensory evidence according to Bayes rule:

$$p(C = 1|x_A, x_V) = \frac{p(x_A, x_V|C = 1)p_{common}}{p(x_A, x_V)}$$

In the case of a single, common, cause (C = 1), the maximum a posteriori probability estimate of the auditory location (S_{AV}) is a reliability-weighted average of the auditory and visual estimates and the prior. In the case of two separate causes (C = 2), the auditory signal location (S_A) is estimated independently from the visual spatial signal (S_A). The final estimate of the auditory location (S_A) is computed using a model averaging strategy, in which the estimates derived under a common cause (C = 1) and separate causes (C = 2) are combined, weighted by the posterior probability of each causal structure ($p(C = 1|x_A, x_V)$, $p(C = 2|x_A, x_V)$).

$$\hat{S}_A = p(C = 1|x_A, x_V)\hat{S}_{AV,C=1} + (1 - p(C = 1|x_A, x_V))\hat{S}_{A,C=2}$$

In our experiment, to assess whether the prior of common cause p_{common} is stronger for audiovisual signals with greater realism, we fit this causal inference model to our participants' data for Experiment 1. We optimized the four model parameters (p_{common}, σ_P^2, σ_A^2, and σ_V^2) for each participant and in each AV Realism condition independently using the Bayesian Adaptive Direct Search (version) toolbox for python (PyBads, v1.0.3) based on the log-likelihood of the true data under the causal inference model and particular parameter settings (Singh and Acerbi, 2024; Acerbi & Ma, 2017).

We initially sampled 5000 parameter settings from a multidimensional uniform distribution of the plausible boundaries for each parameter. Of these, we used the 16 combinations with the highest log likelihood as starting points for the PyBads search algorithm for each participant and each AV Realism condition. The following ranges were used as upper and lower bounds for the fitting each parameter: p_{common}: 0-1, σ_P^2: 0°-100°, σ_A: 0°-100°^{2}, and σ_V^2: 0°-100°. Of the 16 initializations, we obtained the parameter estimate values that resulted in the lowest negative log likelihood for each participant and each AV Realism condition. From this best initialization, we also calculated the coefficient of determination (R^2) for each participant and level of AV Realism as $R^2 = 1 - (SSR/SST)$, where SSR is the residual sum of squares representing the sum of squared differences between the observed and predicted values, and SST is the total sum of squares representing the total variance in the observed data. Mean parameter values and model fit (R^2) are reported in Table 4.2. To evaluate the effect of AV Realism on model parameters, we performed one-way, repeated measures analyses of variance (RM ANOVAs) on each of the parameter values across our three levels of AV Realism and Bonferroni-corrected post-hoc pairwise t-tests for significant RM ANOVAs using the python package Pingouin (Vallat, 2008).

Results

Animated realism enhances crossmodal bias at the smallest AV Disparities.

First, to confirm that there was a basic ventriloquism effect, we examined the GLMM intercept and slope for the Sphere + Tone reference condition. Establishing the pattern of ventriloquism over AV disparities in this condition is important because it serves as the reference data against which ventriloquism in the more realistic object conditions is compared (see Methods). The results confirmed our expectation that crossmodal bias decreased as the disparity between the audio and visual stimuli increased (Sphere + Tone intercept: β = 65.48, *SE* = 3.60, $t(37.55) = 18.17$, $p < 0.0001$; AV Disparity: β = -0.68, *SE* = 0.049, $t(7182.01) = -13.99$, $p < 0.0001$) (Figure 4.3 gray line).

Next, we turned to the primary question of whether stimulus realism changes the strength of ventriloquism. The results showed that the most realistic, Animated Object + Object Sound, condition had a significantly larger intercept value (β = 14.30, *SE* = 2.84, $t(7182.01) = 5.04$, $p < 0.0001$). The intercept of the crossmodal bias was estimated to be more than 14% larger than the reference Sphere + Tone condition. Interestingly, the crossmodal bias also decreased faster across AV disparities than the reference condition (β = -0.31, *SE* = 0.069, $t(7182.02) = -4.50$, $p < 0.0001$). This suggests that the crossmodal bias was stronger when animated objects and their sounds were relatively close together, but this effect dissipated over larger disparities (Figure 4.3, red lines).

In contrast to the animated condition, crossmodal bias in the Static + Congruent condition was not significantly different from the Sphere + Tone condition (intercept: β = 0.037, *SE* = 2.84, $t(7182.01) = 0.013$, $p = 0.990$), and the change in bias over AV disparities was also not significant (β = -0.100, *SE* = 0.069, $t(7182.02) = -1.45$, $p = 0.147$). This was somewhat surprising

because it implies that the presence of a meaningful static object with a congruent sound does not produce a larger crossmodal bias than a simple sphere and tone. Thus, meaningful objects and sounds do not appear sufficient to increase multisensory integration when static.

In order to directly compare the Animated Object + Object Sound and Static Object + Object Sound conditions against each other, the same GLMM model was run again but now with the Static Object + Object Sound condition set as the reference. This revealed that the crossmodal bias in the Animated Object + Object Sound condition was significantly larger at the nearest disparities (β = 14.26, *SE* = 2.85, $t(7182.02) = -4.10$, $p < 0.0001$), and also decreased more quickly as stimulus disparity increased (β = -0.21, *SE* = 0.07, $t(7182.02) = 3.04$, $p = 0.002$). This suggests that there is something unique about an animated object coupled with realistic semantic sounds that enhances multisensory integration at the near disparities.

The GLMM did not directly test crossmodal bias at the specific disparities within our study. Therefore, to verify the finding that crossmodal bias was larger for the Animated Object + Object Sound condition only at the nearest AV disparity within our experimental design, we ran pairwise t-tests on data from the three AV Realism conditions at 15° and 60° of AV disparity (Figure 4.1a). As expected from the GLMM, at 15° of AV Disparity, there was a significant difference between the Animated Object + Object Sound condition (μ = 67.97%, *SD* = 51.09%) and both the Static Object + Object Sound, (μ = 55.26%, *SD* = 61.69%; $t(19) = 3.71$, $p = 0.009$, $BF_{10} = 26.06$), and the Sphere + Tone conditions (μ = 55.95%, *SD* = 62.85%; $t(19) = 3.39$, $p = 0.019$, $BF_{10} = 13.83$). The condition with an animated object and realistic sounds produced significantly larger crossmodal bias at 15° than both the most simplistic and the static, realistic condition. There was no difference between the Sphere + Tone and the Static Object + Object Sound conditions ($t(19) = -0.23$, $p = 1.00$, $BF_{01} = 4.35$).

At 60° of AV Disparity, there was no significant difference in crossmodal bias between the Animated Object + Object Sound (μ = 23.37%, *SD* = 27.94%) and Static Object + Object Sound conditions (μ = 20.54%, *SD* = 24.74%; $t(19) = 1.85$, $p = 0.47$, $BF_{01} = 1.03$), or between the Animated Object + Object Sound and the Sphere + Tone conditions (μ = 26.37%, *SD* = 25.83%; $t(19) = -1.72$, $p = 0.61$, $BF_{01} = 1.24$). Thus, the difference in crossmodal bias shown at small disparities between the animated and other two conditions was not present at the largest disparity. There was, however, a significant difference between the Sphere + Tone and Static Object + Object Sound conditions ($t(19) = -3.65$, $p = 0.01$, $BF_{10} = 23.06$): crossmodal bias was greater for the Sphere + Tone condition. This is interesting because it suggests that meaning may have supported more accurate segregation at 60º of disparity, though more evidence is needed to confirm this.

Animated realism enhances the prior of common cause in the Bayesian causal inference model.

We fit Bayesian causal inference models to our data to investigate how stimulus realism might affect potential mechanisms of multisensory integration. Across AV Realism conditions and participants, the causal inference model accounted for a mean of 29.88% of the variance within our data (see Table 4.1). A one-way RM ANOVA revealed a significant effect of AV Realism on p_{common} ($F(2, 38) = 5.27$, $p = 0.010$, $\eta_p^2 = 0.05$) (Figure 4.4). Bonferroni-corrected post hoc pairwise t-tests showed that p_{common} was greater for the Animated Object + Object Sound condition (μ = 0.75, *SD* = 0.19) than the Static Object + Object Sound condition (μ = 0.63, *SD* = 0.26; $t(19) = 3.07$, $p = 0.02$, $BF_{10} = 7.52$). The difference in p_{common} between the Animated Object + Object Sound and the Sphere + Tone conditions (μ = 0.67, *SD* = 0.19) was not statistically significant, though Bayes Factors provided anecdotal evidence for a greater

p_{common} in the Animated Object + Object Sound condition, ($t(19) = 2.07$, $p = 0.16$, $BF_{10} = 1.35$). There was also no significant difference between p_{common} in the Static Object + Object Sound and Sphere + Tone conditions; $t(19) = -1.17$, $p = 0.77$, $BF_{01} = 2.36$. Additionally, RM ANOVAs did not suggest any significant differences between AV Realism conditions on σ_A^2 ($F(2, 38) = 0.96$, $p = 0.39$, $\eta_p^2 = 0.008$) or σ_V^2 ($F(2, 38) = 0.97$, $p = 0.35$, $\eta_p^2 = 0.02$). There was a significant effect of AV Realism on σ_P^2 ($F(2, 38) = 3.63$, $p = 0.05$, $\eta_p^2 = 0.03$), though none of the post-hoc pairwise comparisons were significant.

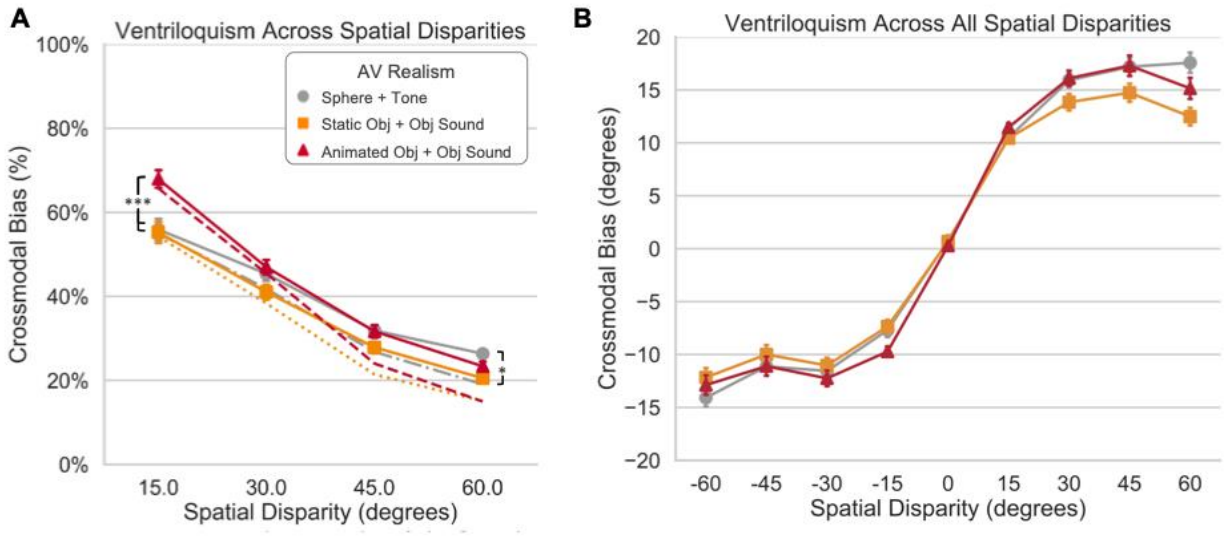


Figure 4.3. Results from Experiment 1. **A**. Crossmodal bias (in percentage) at each level of AV Spatial Disparity, excluding 0° AV Disparity. Solid lines and points represent observed behavior, while broken lines represent causal inference model predictions for each AV Realism condition (dotted: Static + Congruent, dashed: Animated Object + Object Sound, dot-dashed: Sphere + Tone). **B**. Ventriloquism presented as degrees of crossmodal bias at each spatial disparity in the positive or negative direction. All error bars denote standard error of the mean (SEM).

Overall, the significant difference between p_{common} in the Animated Object + Object Sound and Static Object + Object Sound conditions suggests that animated objects produce stronger realism, which affects multisensory processing by enhancing the prior of common cause. It is somewhat surprising that there was not strong evidence for a difference between the Animated Object + Object Sound condition and the Sphere + Tone condition. However, this may

be because participants do not come into the experiment with strong prior expectations about the natural behavior of spheres and tones and may have quickly learned the combination.

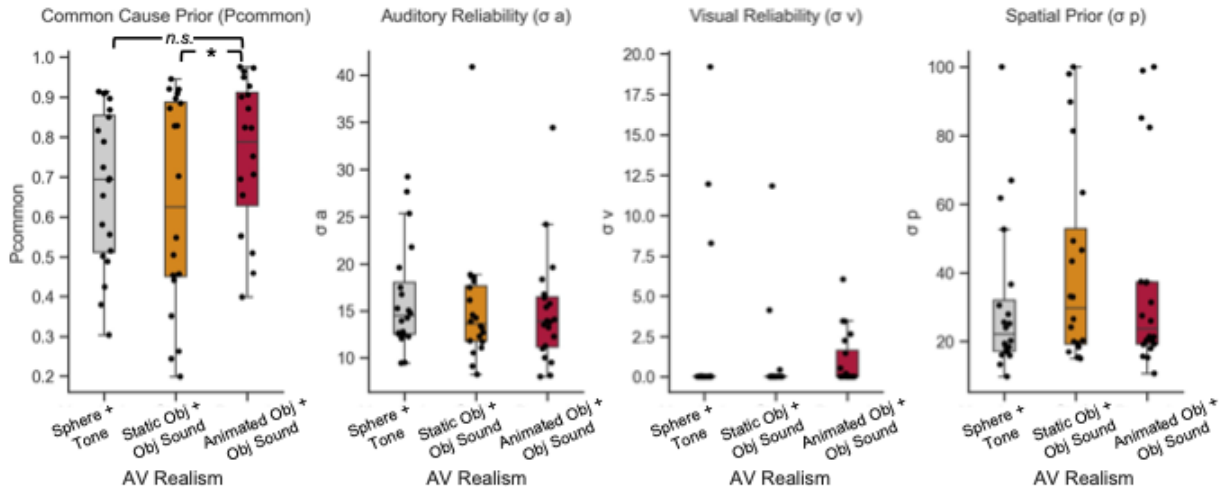


Figure 4.4. Bayesian causal inference model results for Experiment 1. Each plot shows the average of the best model parameters for each participant for each AV Realism condition. On each plot, the central box represents the interquartile range (IQR), marking the middle 50% of the data between the first and third quartiles. The horizontal line inside the box denotes the median value. Whiskers extend up to 1.5 times the IQR from the quartiles, and outliers are indicated by individual points beyond the whiskers.

Table 4.1. Model parameters (across-subjects' mean ± SEM) and fit indices of the computational model from Experiment 1.

Audiovisual Realism	p_{common}	σ_A	σ_V	σ_P	R^2
Sphere + Tone	0.67±0.04	16.24±1.27	1.98±1.15	30.79±5.08	0.20±0.10
Static Object + Object Sound	0.63±0.06	15.28±1.52	0.82±0.62	41.43±6.57	0.40±0.05
Animated Object + Object Sound	0.75±0.04	14.98±1.36	1.00±0.38	37.17±6.51	0.36±0.06

Experiment 2

The purpose of Experiment 2 was to test whether the finding in the Animated Object + Object Sound condition was due to the specific combination of an animated object with a semantically and temporally congruent sound (i.e., a realistic object producing a realistic sound), or due to object animation with any temporally coincident sound. To do this, we conducted a replication experiment but with the Animated Object + Object Sound condition replaced with an Animated Object + Tone condition. The animated objects were identical to Experiment 1, but the

sound was now a temporally synchronous but meaningless tone. The Static Object + Object Sound and Sphere + Tone conditions were identical to those in Experiment 1 to keep the context of the experiment the same for the animated condition. The sample size, hypotheses, and analysis plan for this experiment were pre-registered and can be found on the Open Science Framework (OSF; https://osf.io/zue3h).

Method

Participants. Twenty seven students (19 identified as female, 7 identified as male, and 1 identified as nonbinary, μ_{age} = 19.56 years) from the University of California, Davis, participated in exchange for partial course credit. As in Experiment 1, participants were screened for symptom susceptibility prior to taking part in the study using the Visually Induced Motion Sickness Susceptibility Questionnaire (VIMSSQ; Keshavarz et al., 2019). Three participants were excluded due to poor unisensory sound localization accuracy (see Data Analysis), three participants were excluded because they did not finish the experiment, and one participant was excluded because they did not follow instructions and reported the visual object location instead of the sound location.

Materials.

Apparatus. All details of the apparatus and experimental setup for were identical to Experiment 1.

Stimuli. The stimuli were identical to those in Experiment 1, with the exception of those used in the animated condition. In this experiment, the same short visual animations were now paired with meaningless tones. The tones were presented in synchrony with the animation, just as the congruent sounds were in Experiment 1. For example, a 3350 Hz tone was played at the moment of impact between the drumstick and the drum, analogous to the onset of the snare drum

sound in Experiment 1. Thus, high temporal correspondence was maintained between the sound and the visual animation, without there being a semantic relationship between them. The 1045 Hz tone was paired with the scissors and the toaster, and the 3350 Hz tone was paired with the drum and the wine bottle. Each tone was paired with two objects but participants only saw one of the objects paired with each tone (see Experimental Design below). The two new tones at 1045 Hz and 3350 Hz were different from that used in the Sphere + Tone condition (1700 Hz tone). The peak amplitudes of all sounds were normalized using Audacity software (Audacity Team, 2021).

Experimental Design. The difference between this study and Experiment 1 was that the animated objects were now paired with meaningless tones instead of congruent sounds. AV Disparity was manipulated in the same manner as Experiment 1, by presenting participants with synchronous audiovisual stimulus pairs presented either at the same or disparate locations. The auditory and visual stimulus locations were sampled independently from five possible locations (angles -30º, -15º, 0º, 15º, and 30º) at a distance of 5 m along the azimuth axis. Thus, stimuli could be presented at one of five levels of AV Disparity (none/0°, 15º, 30º, 45º, or 60º). An equal number of each possible location combinations was used for each AV Disparity.

AV Realism was also manipulated similarly to Experiment 1, with the exception that the animated objects were now paired with meaningless tones. The audiovisual stimulus pairs on each trial were either the red sphere and the 1700 hz tone (Sphere + Tone), a realistic static object and its respective sounds (Static Object + Object Sound), or a realistic animated object and a meaningless tone (Animated Object + Tone). Because the realistic objects were only paired with semantically congruent sounds in one condition, each participant was only shown two realistic objects in each condition (two in the Static Object + Object Sound and a different two in

the Animated Object + Tone condition) to avoid cross-condition interference. Each realistic visual object appeared in only one AV Realism condition and they were counterbalanced across participants in two versions. Version 1 included the drum and scissors in the Static+ Congruent sound condition and the wine bottle and toaster in the Animated Object + Tone condition, and version 2 had the opposite.

Procedure. On each trial, the visual stimulus was presented for 1000 ms, and the sound was played in its entirety (100-150 ms). On trials with a sphere or a static object, the sound was synchronized with the onset of the visual stimulus. On trials with an animated object, the tone was synchronized with the animation (e.g., at the moment of impact of the drumstick on the drum) (Figure 4.1). After the stimuli were presented, nine response options were displayed as gray spheres, located along the azimuth axis, spanning from -30° to 30° degrees at 7.5° intervals (Figure 4.2).

The response instructions and procedures were identical to that of Experiment 1. In short, participants were instructed to report the perceived location of the sound by pointing a virtual laser. Participants had a maximum of 2000 ms to respond, and a 1000 ms inter-trial interval (ITI) followed each response. Participants were informed that the sounds and visual objects could originate from the same or different locations, but were not informed about the frequency of same versus different location trials, nor that the sound could only originate from five of the nine response options. Participants completed 550 trials, split into three blocks to allow for breaks outside of the headset.

There were 35 trials per condition x 3 (AV Realism: Sphere + Tone, Static Object + Object Sound, Animated+ Tone) x 5 (AV Disparity: 0°, 15º, 30º, 45º, 60º) for a total of 525 trials. An additional 25 unisensory visual trials were used as attention checks, in which visual

stimuli sampled from all visual object types were presented with no sound, and participants were instructed to report the visual location on these trials. Each of the five visual stimuli and locations were sampled equally for these visual attention check trials. Trial order was randomized for each participant. Ten practice trials were completed prior to the main task, and these were repeated as needed for participants to understand the task. Practice trials and unisensory visual trials were not included in the analyses.

Prior to the main experiment, participants completed the same unisensory visual and auditory localization task as in Experiment 1 to ensure that they could reliably localize the visual and auditory stimuli individually. In total, the study took approximately one hour to complete.

Data Analysis.

Unisensory localization performance. As in Experiment 1, we assessed performance on the unisensory localization trials to ensure that included participants could reliably localize both auditory and visual stimuli within our VR experimental setup. Overall accuracy on the visual localization task was very high (visual sphere mean = 99.00% correct, SD = 9.95%; visual static mean = 97.04% correct, SD = 16.98; visual animated object mean = 96.55% correct, SD = 18.29%). A repeated measures analysis of variance (ANOVA) did not show any significant differences in localization accuracy across these three visual object types. As expected, accuracy on the auditory localization task was lower overall (1700 Hz tone mean = 30.00% correct, SD = 45.94%; object sounds mean = 34.75% correct, SD = 47.68%; 3350 Hz and 1045 Hz tone mean = 37.00% correct, SD = 48.34). A repeated measures ANOVA did not show any significant differences in localization accuracy across these three sound types.

We also calculated the Pearson correlation between participant responses and the true signal source. This showed that participants were able to localize both unisensory visual and

auditory stimuli reliably, as indicated by positive Pearson correlations between participants' location responses and the stimulus locations: unisensory realistic object sounds (across participants mean +/– SEM: 0.89 +/– 0.01), the 3350 Hz and 1045 Hz sounds (across participants mean +/– SEM: 0.88 +/– 0.02), and for the 1700 Hz sound (across participants mean +/– SEM: 0.85 +/– 0.02). The correlations were also significant for the static visual objects (across participants mean +/– SEM: 0.98 +/– 0.02), animated visual objects (across participants mean +/– SEM: 0.98 +/– 0.02) and the sphere (across participants mean +/– SEM: 1.00 +/– 0.002; see Supplemental Materials). These results demonstrate that, as in Experiment 1, participants could reliably locate unisensory auditory and visual stimuli within our VR experimental setup.

Mixed effects modeling of crossmodal bias. We tested our main predictions for this experiment using a GLMM. The analysis procedure was the same as Experiment 1, including the likelihood ratio test between the full and reduced models, and analysis of the full model output. The full model included fixed effects of AV Disparity (15º, 30º, 45º, 60º) and AV Realism (Sphere + Tone, Static Object + Object Sound, Animated Object + Tone), their interaction, and a random effect of participant. The reduced model was the same, except excluding of the fixed effects of AV Realism. The procedures for comparing the models was the same as in Experiment 1. The likelihood ratio test indicated that the full model including AV Realism as a predictor provided a better fit for the data than a model without it (reduced model), χ^2 (8) = 82.26, $p <$ 0.0001.

As in Experiment 1, AV Realism was dummy-coded with Sphere + Tone as the reference level, representing the baseline level realism. In addition to reporting beta coefficients for the model, we also had specific predictions for differences in crossmodal bias between AV Realism conditions for the 15° and 60° AV Disparity levels, which we tested with Bonferroni-corrected

pairwise t-tests, as in Experiment 1. Specifically, we performed pairwise t-tests between the three AV Realism conditions (Sphere + Tone, Static Object + Object Sound, Animated Object + Tone) at 15°and 60º of AV Disparity.

Cross-experiment mixed effects modeling of crossmodal bias. In addition to our pre-registered analyses, to directly compare the effects of animated stimuli paired with congruent sounds versus tones on crossmodal bias, we ran a GLMM analysis on the Animated Object + Object Sound condition from Experiment 1 and the Animated Object + Tone condition from Experiment 2 between subjects. For this analysis, we used the subset of data from each experiment corresponding to these conditions. As in the other GLMM analyses, we treated AV Disparity as a continuous variable, and AV Realism as a factor with two levels (Animated Object + Object Sound, Animated Object + Tone). The full model included fixed effects of AV Disparity and AV Realism, the interaction between them, and a random effect of participant. To confirm that there was an overall effect of AV Realism on crossmodal bias, we performed a likelihood ratio test between the full model compared to a reduced model with only AV Disparity as a predictor. The likelihood ratio test indicated that the full model including AV Realism as a predictor provided a better fit for the data than a model without it (reduced model), χ^2 (6) = 17.76, p = 0.00014. This confirms that AV Realism affects the strength of the crossmodal bias within these experimental conditions.

As with the main analyses for Experiments 1 and 2, we also conducted t-tests between AV Realism conditions at two AV Disparity levels of interest: 15° and 60°. This allowed us to test whether there were specific differences in crossmodal bias between the two animated object conditions at the smallest and largest disparities in the experiments.

Bayesian Causal Inference model. We also fit a Bayesian causal inference model to the data in each condition and for each participant to keep the analysis plan consistent with Experiment 1. However, we did not expect the model parameters to differ between conditions of our experiment, so we have reported these exploratory results in the supplemental materials. As expected, we did not find differences between any model parameters in this study (see Appendix 1 of supplemental materials).

Results

Animated objects do not increase crossmodal bias when paired with meaningless tones.

As in Experiment 1, we first confirmed the presence of a basic ventriloquism effect in the Sphere + Tone reference condition. The GLMM again showed a significant intercept (β = 67.06, *SE* = 3.45, $t(35.18) = 19.421$, $p < 0.0001$) and an estimated decrease in crossmodal bias by 0.67% with each degree of AV disparity (β = -0.67, *SE* = 0.045, $t(8399) = -15.11$, $p < 0.0001$). These results replicate the finding from Experiment 1 and expectations from the literature that crossmodal bias is present with simple stimuli but its strength decreases as stimulus disparity increases.

Next, we looked at how the animated object condition, which was paired with a meaningless tone, compared to the reference Sphere + Tone condition. The results showed that the crossmodal bias in the Animated Object + Tone condition was significantly smaller than the Sphere + Tone condition (β = 5.59, *SE* = 2.60, $t(8399) = -2.149$, $p = 0.0317$). The change in crossmodal bias over AV disparities was not statistically significant (β = -0.088, *SE* = 0.063, $t(8399) = -1.39$, $p = 0.166$). This means that an animated stimulus paired with a meaningless tone actually elicited a weaker crossmodal bias than the simplistic sphere paired with a tone, and there

was no difference in sensitivity to changes in crossmodal bias over disparities. This result is opposite to that of Experiment 1 in which the Animated Object + Object Sound condition showed an increase in crossmodal bias relative to the Sphere + Tone condition at the nearest disparities and a faster drop off across disparities. This shows that the effect of animated realism on strengthening ventriloquism was eliminated, or even reversed, in the absence of a meaningful sound.

The Static Object + Object Sound condition did not significantly alter crossmodal bias relative to the Sphere + Tone condition (β = 2.78, *SE* = 2.60, $t(8399) = 1.07$, $p = 0.29$), but it did decrease more rapidly over AV disparities (β = -0.213, *SE* = 0.063, $t(8399) = -3.368$, $p < 0.001$). The latter result is in line with the t-test results from Experiment 1, suggesting that there was less crossmodal bias at large stimulus disparities for meaningful static objects, perhaps because meaning supports more accurate segregation and reduces the illusion of ventriloquism, or because auditory precision was numerically different across these conditions.

As in Experiment 1, we also re-referenced the model to assess differences in crossmodal bias between the Static Object + Object Sound and the Animated Object + Tone conditions by setting the reference level of AV Realism to Static Object + Object Sound. In this case, crossmodal bias was significantly stronger in the Static condition than the Animated one (β = -8.36, *SE* = 2.61, $t(8389) = 3.201$, $p = 0.001$). The crossmodal bias also decreased faster over AV disparities in the Static Object + Object Sound condition (β = 0.13, *SE* = 0.06, $t(8389) = -1.98$, $p = 0.048$). These data show that the animated object produced weaker ventriloquism when paired with a meaningless tone than a static object paired with a semantically congruent sound. A meaningless tone actually interfered with multisensory integration, highlighting the importance of audiovisual realism for ventriloquism.

To test for differences in AV Realism conditions at the smallest and largest stimulus disparities specifically, we ran pairwise t-tests between the three AV Realism conditions at 15° and 60° of disparity (Figure 4.3a). At 15° of AV Disparity, there were no significant differences in crossmodal bias between the Animated Object + Tone (μ = 53.38%, *SD* = 60.96%) and the Sphere + Tone condition (μ = 59.69%, *SD* = 58.38%; $t(19) = -1.92$, $p = 0.21$, $BF_{01} = 0.93$) nor the Static Object + Object Sound conditions (μ = 60.12%, *SD* = 58.81% ; $t(19) = -1.88$, p = 0.23, $BF_{01} = 0.99$). Bayes Factors showed moderate evidence for the null hypothesis between the Static Object + Object Sound condition and the Sphere + Tone condition ($t(19) = 0.11$, p = 1.00, $BF_{01} = 4.27$), though this was not statistically significant. These results are consistent with our hypothesis from Experiment 1 that realistic, animated objects only elicit greater ventriloquism at small distances when there is semantic correspondence between the visual object and the sound. At 60° of AV Disparity, the crossmodal bias in the Sphere + Tone condition (μ = 29.65%, *SD* = 26.13%) was significantly larger than in the Animated Object + Tone condition (μ = 19.63%, *SD* = 23.13%; $t(19) = 4.17$, $p = 0.0016$, $BF_{10} = 65.02$) and in the Static Object + Object Sound condition (μ = 21.13%, *SD* = 26.37%; $t(19) = 5.31$, $p = 0.00012$, $BF_{10} = 637.19$). There was no significant difference in crossmodal bias between the Animated Object + Tone and the Static Object + Object Sound conditions at 60° ($t(19) = -0.82$, $p = 1.00$, $BF_{01} = 3.18$). These results are consistent with those from Experiment 1 showing weaker ventriloquism for realistic visual objects at the largest disparity than the simple Sphere + Tone stimuli.

Comparison of animated objects across experiments shows larger crossmodal bias when paired with congruent sounds than tones.

The results from Experiments 1 and 2 suggest that animated realism produced greater sensitivity to ventriloquism across VA disparities when paired with a semantically and

temporally congruent sound. Interestingly, the opposite was true when the animated object was paired with a temporally congruent but meaningless tone. This implies that realistic visual animation alone is insufficient to enhance multisensory integration. Although the two conditions differed from the same Sphere + Tone reference, they were never directly compared. To formally test for a difference between the two animated conditions, we ran an additional GLMM with only the two animated conditions from Experiments 1 and 2.

The results of the GLMM revealed that compared to the Animated Object + Tone condition, the Animated Object + Object Sound condition had a stronger crossmodal bias at the smallest estimated disparities (β = 18.33, SE = 5.00, t(66.57) = 3.67, p < 0.0005) and decreased significantly faster in the crossmodal bias over AV disparities (β = 0.23, SE = 0.06, t(5147.99) = 3.61, p = 0.0003). This suggests that multisensory integration was more sensitive in the

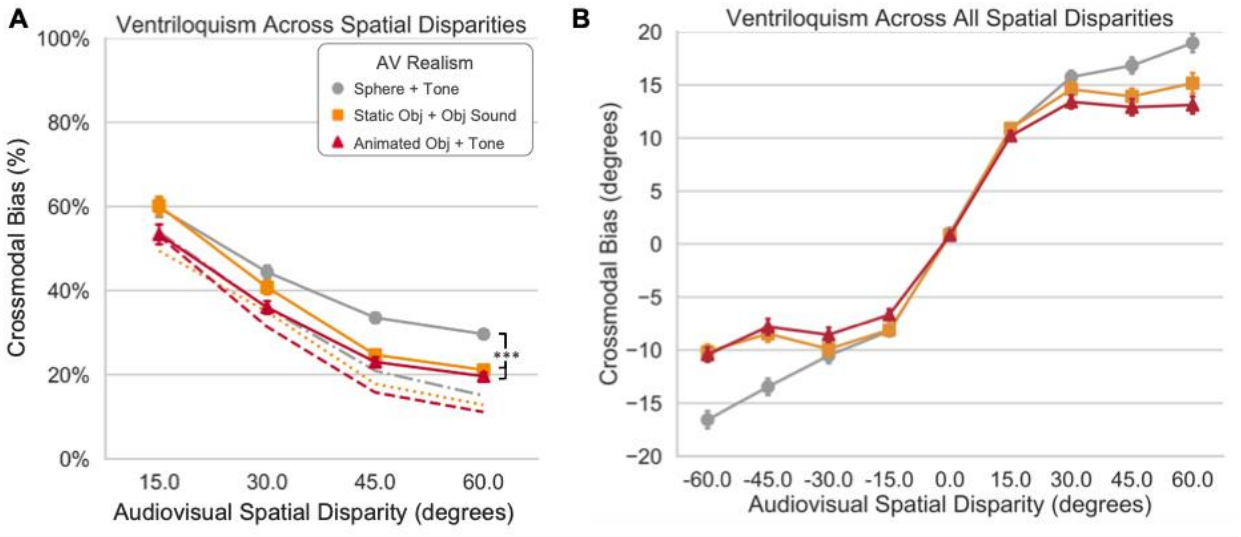


Figure 4.5. Results from Experiment 2. **A**. Crossmodal bias (in percentage) at each level of AV Spatial Disparity, excluding 0°. Solid lines and points represent observed behavior, while broken lines represent causal inference model predictions for each AV Realism condition (dotted: Static Object + Object Sound, dashed: Animated Object + Tone, dot-dashed: Sphere + Tone). **B**. Ventriloquism presented as degrees of crossmodal bias at each spatial disparity in the positive or negative direction. All error bars denote standard error of the mean (SEM).

Animated Object + Object Sound condition, with stronger ventriloquism at near AV disparities and a quicker drop-off as the distance between the audio and visual stimuli increased (Figure 4.6).

As with our approach to the Experiment 1 and 2 analyses, we also tested differences in AV Realism conditions at the smallest and largest stimulus disparities directly using pairwise t-tests (Figure 4.6). At 15° of AV Disparity, there was significantly stronger crossmodal bias in the Animated Object + Object Sound condition (μ = 67.99%, *SD* = 19.92%) than the Animated Object + Tone condition, (μ = 53.40%, *SD* = 20.43%; $t(38) = 2.29$, $p = 0.027$). This replicates the GLMM result and supports the idea that animated, realistic objects elicit greater crossmodal bias when paired with a meaningful sound than when paired with a meaningless tone, even when the tone corresponds temporally to the animation. There was not a significant difference between the Animated Object + Object Sound (μ = 23.26%, *SD* = 16.82%) and Animated Object + Tone (μ = 19.62%, *SD* = 13.31%) conditions at 60° of AV Disparity ($t(38) = 0.76$, $p = 0.45$). Thus, although there is more crossmodal bias in the Animated Object + Object Sound condition at small disparities, the amount of bias is similarly low at larger AV disparities.

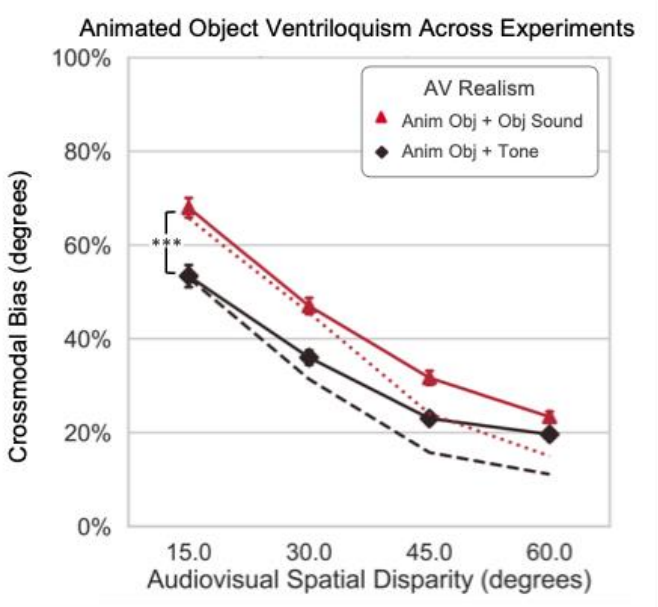


Figure 4.6. Crossmodal bias (in percentage) at each level of AV Spatial Disparity, excluding 0°. All error bars denote standard error of the mean (SEM).

Discussion

In the present work, we investigated the effects of stimulus realism on audiovisual integration within a spatial ventriloquism paradigm presented in virtual reality. We tested the strength of ventriloquism using simplistic stimuli, similar to those commonly used in previous studies, as well as realistic static and animated objects, which allowed us to dissociate the effects of meaningful realistic relationships between audiovisual stimuli and physical movement. In Experiment 1, we found that realistic animated objects and their congruent sounds led to increased ventriloquism at the smallest spatial disparities. Bayesian causal inference modeling suggested that this may be because animated motion and semantic information, in combination, increase the belief that the stimuli originated from the same source. In Experiment 2, we found that the animated objects no longer led to greater ventriloquism when paired with a meaningless tone. This suggests that the findings from Experiment 1 were due to a combination of motion and audiovisual semantic congruence, rather than solely due to animated motion. Together, our studies provide evidence that stimulus realism can increase spatial ventriloquism when there is both dynamic motion and meaningful correspondence between audiovisual stimuli.

Our work provided evidence that meaningful correspondence between realistic auditory and visual stimuli is not sufficient to increase spatial ventriloquism on its own. In Experiment 1, we expected that the Static Object + Object Sound stimuli would elicit stronger ventriloquism than the Sphere + Tone given the potential for prior knowledge of objects and the sounds they produce to increase the belief that the stimuli originated from the same source. However, we did not find this to be the case. When the same visual stimuli were animated to reflect the time at which their characteristic sounds were produced (Animated Object + Object Sound), they did elicit stronger ventriloquism at near spatial disparities compared to both the Sphere + Tone and the Static Object + Object Sound combinations. This suggests that meaning alone was not

enough to functionally increase the belief that stimuli come from the same source and impact integration across space. In the context of Bayesian causal inference models, this was reflected by a larger prior belief of common cause for the Animated Object + Object Sound condition than the Static Object + Object Sound condition. However, neither condition was significantly different in the prior of common cause from the simplistic Sphere + Tone condition, so more research is necessary to better characterize the mechanisms underlying this effect.

Results from Experiment 2 further refine our conclusions by showing that highly correlated, animated audiovisual stimuli that lack semantic correspondence do not elicit stronger ventriloquism than simplistic stimuli. In Experiment 2, we maintained the temporal alignment between the realistic animated visual objects and the meaningless tones by synchronizing the tones with the visual animations. If the temporal correspondence alone was enough to elicit the stronger ventriloquism found in Experiment 1 for the animated object + congruent sound condition, then we should have seen the same result when the animated objects were paired with meaningless tones. This was not the case, suggesting that semantic correspondence is a necessary component of the animation dynamics that increase the ventriloquist effect. This suggests that the results of the studies reported in Jackson (1953) may have been due to the combination of meaningful correspondence between the audiovisual stimuli and motion, rather than meaning or stimulus realism alone.

Our findings are distinct from similar studies that have also used realistic stimuli in a spatial ventriloquism paradigm. Huisman et al. (2022) did not find stronger ventriloquism for their realistic condition compared to less realistic stimuli. However, their realistic stimulus pairing (handball + bounce sound) may not have activated strong enough existing semantic information to increase ventriloquism compared to a simpler stimulus condition that could be

quickly learned. Similarly, Radeau & Bertelson (1987, 1988) did not find any differences in localization recalibration following exposure to realistic audiovisual stimuli (i.e., the ventriloquism aftereffect) compared to animated yet meaningless audiovisual stimuli. This could be because dynamic motion and meaningful correspondence are more effective for the ventriloquist effect than the aftereffect. As noted by Bruns (2019), the perception of unity based on meaning may affect in-the-moment multisensory integration without impacting the type of learning that leads to the ventriloquist aftereffect. Another potentially important difference between our experimental design and that of Radeau & Bertelson (1987, 1988), was that their control condition included meaningless visual stimuli (light flashes) paired with meaningful sounds (drums, speech), whereas we used meaningful visual objects and meaningless tones. It is possible that they saw comparable recalibration between conditions because there is an asymmetry in the effects of semantic content within auditory versus visual information on increasing integration.

Across both of our experiments, we found stronger ventriloquism at the largest disparity (60°) for the simplistic Sphere + Tone stimuli compared to stimulus pairs with realistic objects. This was only significant for the Static Object + Object Sound condition in Experiment 1, but it was true for both the Static Object + Object Sound and the Animated Object + Tone conditions in Experiment 2. Interestingly, this decrease in integration at the farther disparities could be because meaning might also be useful in correctly deciding that stimuli *do not* belong to a single source. In another spatial ventriloquism study, Wallace et al. (Wallace et al., 2004) found either no bias, or a repelling effect rather than ventriloquism for stimuli that had been judged as originating from two distinct sources, and similar repelling effects have been observed in other studies (e.g., Körding et al., 2007; Rohe & Noppeney, 2015). It is possible that, within our study,

the meaning within object stimuli increased this repelling effect relative to the neutral sphere + tone. Future studies could test this hypothesis explicitly. Overall, this finding suggests that there may be multiple mechanisms by which meaningful information contributes to multisensory integration.

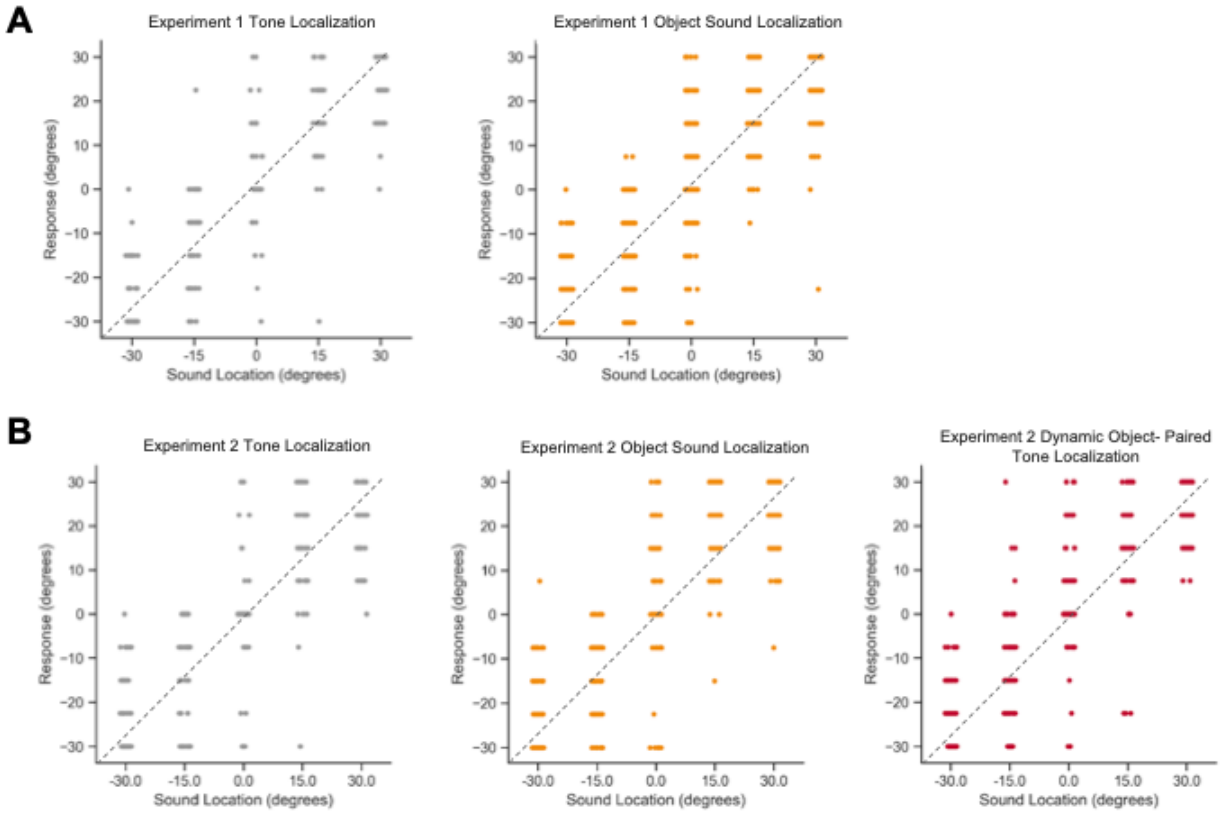
In conclusion, the present results provide evidence that spatial ventriloquism is stronger across small spatial disparities for meaningful, animated audiovisual stimuli than for the types of simplistic stimuli frequently used in studies of multisensory integration. In contrast, we also found that ventriloquism was weaker for stimulus pairs with meaningful stimuli at large disparities compared to simplistic stimuli. These distinct effects across stimulus pairings and spatial disparities elucidate a nuanced impact of realistic stimuli on ventriloquism. This underscores the need to further investigations into the interaction of top-down and stimulus driven influences on multisensory integration in complex real-world environments. The development of tools like virtual and augmented reality will aid in our ability to investigate many aspects of naturalistic multisensory processing by allowing for tight experimental control and manipulation of three-dimensional crossmodal stimuli. Overall, the present work takes a step towards understanding multisensory integration and its mechanisms with more naturalistic stimuli, by showing that animated realism has a nonlinear impact on multisensory integration.

Chapter 4 Supplemental Materials

Motion Sickness Questionnaire.

To screen for susceptibility to motion and cyber sickness prior to participation in both experiments, we used the Visually Induced Motion Sickness Susceptibility Questionnaire (VIMSSQ; Keshavartz et al., 2019). This questionnaire asks about frequency of use of visual devices such as 2D/3D movies, smartphones, head-mounted displays, video games, and simulators, and feelings of nausea, headache, fatigue, dizziness, and eye strain during past use of each of these devices during adulthood. There were five Likert-scale response options ranging from “often” to “never”. Any participants that reported feeling any of these symptoms “often,” were ineligible for the study, though no participants reached this criterion in our study. The additional open-ended question was also used to assess any additional motion sickness concerns on a case-by-case basis.

Unisensory Sound Localization Accuracy



Supplemental Figure 4.1. This figure depicts the localization reliability for sounds presented within each AV Realism condition for Experiment 1 (A) and Experiment 2 (B). Each figure shows the correlation between the true sound location (x-axis) and the reported sound location (y-axis).

Experiment 2 Supplemental Results.

Bayesian Causal Inference Modeling. As an exploratory analysis, in line with Experiment 1, we also fit Bayesian causal inference models to our data to investigate potential multisensory integration mechanisms affected by stimulus pairings in Experiment 2. Across AV Realism conditions and participants, the causal inference model accounted for a mean of 45.22% of the variance within our data (Supplementary Table 4.1). A one-way RM ANOVA revealed a significant effect of AV Realism on σ_P^2, $F(2, 38) = 4.24$, $p = 0.04$, $\eta_p^2 = 0.02$. However, none of the post-hoc pairwise comparisons were significant. Additionally, RM ANOVAs did not suggest any significant differences between AV Realism conditions on p_{common} ($F(2, 38) = 0.93$, $p = 0.40$,

η_p^2 = 0.01), σ_A^2 (F(2, 38) = 2.31, p = 0.11, η_p^2 = 0.010), or σ_V^2 (F(2, 38) = 0.84, p = 0.37, η_p^2 = 0.03). These null results suggest that none of the Bayesian causal inference model parameters were significantly impacted by object realism or dynamic motion within this experiment. This is not surprising given that, in Experiment 1, the main result from the causal inference model analysis was that p_{common} was impacted by the Animated Object + Object Sound condition, likely because of the strength of the ventriloquist effect for this condition at 15° AV Disparity. Because we did not observe this stronger ventriloquism for the dynamic objects in the absence of congruent sounds for Experiment 2, it follows that there would not be a similarly strengthened prior belief that the stimuli in this condition belong to a common source.

Supplementary Table 4.1. Model parameters (across-subjects' mean ± SEM) and fit indices of the computational model from Experiment 2.

Audiovisual Realism	p_{common}	σ_A	σ_V	σ_P	R^2
Sphere + Tone	0.63±0.04	14.53±1.02	0.84±0.58	27.95±6.56	0.40±0.05
Static Object + Object Sound	0.69±0.05	13.66±1.48	3.85±2.84	38.31±8.12	0.46±0.06
Animated Object + Tone	0.64±0.05	13.22±1.24	1.76±0.81	38.13±7.64	0.50±0.06

References

Acerbi, L. & Ma, W. J. (2017). Practical Bayesian Optimization for Model Fitting with Bayesian Adaptive Direct Search. In *Advances in Neural Information Processing Systems 31*: 8222-8232.

Alais, D., & Burr, D. (2004). The Ventriloquist Effect Results from Near-Optimal Bimodal Integration. *Current Biology*, *14*(3), 257–262. https://doi.org/10.1016/j.cub.2004.01.029

Audacity Team (2021). Audacity(R): Free Audio Editor and Recorder [Computer application]. Version 3.0.0 retrieved March 20th, 2022 from https://audacityteam.org/.

Bates D, Mächler M, Bolker B, Walker S (2015). "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software*, **67**(1), 1–48. doi:10.18637/jss.v067.i01.

Biederman, I. (1972). Perceiving real-world scenes. Science, 177(4043), 77–80.

Bruns, P. (2019). The Ventriloquist Illusion as a Tool to Study Multisensory Processing: An Update. *Frontiers in Integrative Neuroscience*, *13*, 51. https://doi.org/10.3389/fnint.2019.00051

Chen, Y.-C., & Spence, C. (2010). When hearing the bark helps to identify the dog: Semantically-congruent sounds modulate the identification of masked pictures. *Cognition*, *114*(3), 389–404. https://doi.org/10.1016/j.cognition.2009.10.012

Chen, Y.-C., & Spence, C. (2017). Assessing the Role of the 'Unity Assumption' on Multisensory Integration: A Review. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00445

Delong, P., Noppeney, U. Semantic and spatial congruency mould audiovisual integration depending on perceptual awareness. *Sci Rep* **11**, 10832 (2021). https://doi.org/10.1038/s41598-021-90183-w

Jackson, C. V. (1953). Visual Factors in Auditory Localization. *Quarterly Journal of Experimental Psychology*, *5*(2), 52–65. https://doi.org/10.1080/17470215308416626

Kanaya, S., & Yokosawa, K. (2011). Perceptual congruency of audio-visual speech affects ventriloquism with bilateral visual stimuli. *Psychonomic Bulletin & Review*, *18*(1), 123–128. https://doi.org/10.3758/s13423-010-0027-z

Kayser, C., & Shams, L. (2015). Multisensory Causal Inference in the Brain. *PLOS Biology*, *13*(2), e1002075. https://doi.org/10.1371/journal.pbio.1002075

Keshavarz, B., Saryazdi, R., Campos, J. L., & Golding, J. F. (2019). Introducing the VIMSSQ: Measuring susceptibility to visually induced motion sickness. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *63*(1), 2267–2271. https://doi.org/10.1177/1071181319631216

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719. https://doi.org/10.1016/j.tins.2004.10.007

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal Inference in Multisensory Perception. *PLoS ONE*, *2*(9), e943. https://doi.org/10.1371/journal.pone.0000943

Matusz, P. J., Wallace, M. T., & Murray, M. M. (2017). A multisensory perspective on object memory. *Neuropsychologia*, *105*, 243–252. https://doi.org/10.1016/j.neuropsychologia.2017.04.008

Noppeney, U. (2021). Perceptual Inference, Learning, and Attention in a Multisensory World. *Annual Review of Neuroscience*, *44*(1), 449–473. https://doi.org/10.1146/annurev-neuro-100120-085519

Parise, C., & Spence, C. (2008). Synesthetic congruency modulates the temporal ventriloquism effect. *Neuroscience letters*, *442*(3), 257–261. https://doi.org/10.1016/j.neulet.2008.07.010

Parise, C. V., & Spence, C. (2009). 'When birds of a feather flock together': synesthetic correspondences modulate audiovisual integration in non-synesthetes. *PloS one*, *4*(5), e5664. https://doi.org/10.1371/journal.pone.0005664

Park, H., & Kayser, C. (2022). The context of experienced sensory discrepancies shapes multisensory integration and recalibration differently. *Cognition*, *225*, 105092. https://doi.org/10.1016/j.cognition.2022.105092

Radeau, M., & Bertelson, P. (1974). The After-Effects of Ventriloquism. *Quarterly Journal of Experimental Psychology*, *26*(1), 63–71. https://doi.org/10.1080/14640747408400388

Radeau, M., & Bertelson, P. (1977). Adaptation to auditory-visual discordance and ventriloquism in semirealistic situations. *Perception & Psychophysics*, *22*(2), 137–146. https://doi.org/10.3758/BF03198746

Radeau, M., & Bertelson, P. (1978). Cognitive factors and adaptation to auditory-visual discordance. *Perception & Psychophysics*, *23*(4), 341–343. https://doi.org/10.3758/BF03199719

Recanzone, G. H. (1998). Rapidly induced auditory plasticity: The ventriloquism aftereffect. *Proceedings of the National Academy of Sciences*, *95*(3), 869–875. https://doi.org/10.1073/pnas.95.3.869

Recanzone, G. H. (2009). Interactions of auditory and visual stimuli in space and time. *Hearing Research*, *258*(1–2), 89–99. https://doi.org/10.1016/j.heares.2009.04.009

Rohe, T., Ehlis, AC., & Noppeney, U. (2019). The neural dynamics of hierarchical Bayesian causal inference in multisensory perception. *Nature Communications*, *10*(1), 1907. https://doi.org/10.1038/s41467-019-09664-2

Rohe, T., & Noppeney, U. (2015). Sensory reliability shapes perceptual inference via two mechanisms. *Journal of Vision*, *15*(5), 22. https://doi.org/10.1167/15.5.22

Rohe, T., & Noppeney, U. (2016). Distinct Computational Principles Govern Multisensory Integration in Primary Sensory and Association Cortices. *Current Biology*, *26*(4), 509–514. https://doi.org/10.1016/j.cub.2015.12.056

Singh, S. G. & Acerbi, L. (2024). PyBADS: Fast and robust black-box optimization in Python. *Journal of Open Source Software*, 9(94), 5694, https://doi.org/10.21105/joss.05694

Slutsky, D. A., & Recanzone, G. H. (2001). Temporal and spatial dependency of the ventriloquism effect: *Neuroreport*, *12*(1), 7–10. https://doi.org/10.1097/00001756-200101220-00009

Van Wanrooij, M. M., Bremen, P., & John Van Opstal, A. (2010). Acquired prior knowledge modulates audiovisual integration. *European Journal of Neuroscience*, *31*(10), 1763–1771. https://doi.org/10.1111/j.1460-9568.2010.07198.x

Wallace, M. T., Roberson, G. E., Hairston, W. D., Stein, B. E., Vaughan, J. W., & Schirillo, J. A. (2004). Unifying multisensory signals across time and space. *Experimental Brain Research*, *158*(2). https://doi.org/10.1007/s00221-004-1899-9

Warren, D. H., Welch, R. B., & McCarthy, T. J. (1981). The role of visual-auditory "compellingness" in the ventriloquism effect: Implications for transitivity among the spatial senses. *Perception & Psychophysics*, *30*(6), 557–564. https://doi.org/10.3758/BF03202010

Williams, J. R., Markov, Y. A., Tiurina, N. A., & Störmer, V. S. (2022). What You See Is What You Hear: Sounds Alter the Contents of Visual Perception. *Psychological Science*, 095679762211213. https://doi.org/10.1177/09567976221121348

Wozny, D. R., Beierholm, U. R., & Shams, L. (2010). Probability Matching as a Computational Strategy Used in Perception. *PLoS Computational Biology*, *6*(8), e1000871. https://doi.org/10.1371/journal.pcbi.1000871

Chapter 5: General Discussion

Collectively, the work presented in this dissertation advances our understanding of how naturalistic multisensory stimuli are integrated, perceived, and remembered. Efforts to integrate methodologies and findings from traditionally disparate fields can be a challenge, as can the development of materials for experimental designs with novel technologies like virtual reality. However, I would argue that these efforts are worthwhile if they allow us to better explain and predict real-world behaviors and/or if the resulting findings can be applied to develop novel methods for improving technology or education. Below, I discuss the opportunities and challenges that new technologies afford for cognitive research, as well as some emerging ways that these methods are applied for the development of new technology.

Novel research methods and challenges

More than ever, there is an abundance of technology that allows for more naturalistic research methods within the cognitive sciences. The two primary limitations that I encountered during the development of the studies presented in this dissertation included: 1) sourcing good quality, realistic, and artistically cohesive visual and auditory stimuli, and 2) developing VR software for presenting experiments and recording data using tools primarily developed for video game design. However, throughout my time in graduate school, there has been a rapid emergence of novel technology that will undoubtedly reduce these challenges.

There are a number of standardized stimulus sets that have been indispensable in many studies of visual and multisensory processing (e.g., Brodeur et al., 2010; Snodgrass & Vanderwart, 1980; Schneider et al., 2008). However, many stimuli within these can be limited in their realism, relation to stimuli presented in other modalities, artistic similarity and salience to one another, and quantity. Additional sources of stimuli often include Google Images or, as was

the case almost all of the stimuli used in the studies for this dissertation, the Unity Asset Store, though it can be challenging to fabricate cohesive sets of stimuli that are similar in style and salience, especially for studies that require many trials. Generative artificial intelligence (AI) has led to the development of many free and low-cost models for generating stimuli including text-to-image visual objects (e.g., Adobe Firefly, Google's ImageFX) and sounds (e.g., Meta's AudioGen). These tools will undoubtedly be useful for generating stimuli that can be adapted for a number of research areas, especially as they can allow for the use of reference images for creating variations or edits of a single stimulus, as well as embedding visual object stimuli seamlessly into scenes. While the quality of these generated images can vary, models are rapidly improving, and can produce high-quality generated artwork with effective prompt engineering. Text-to-sound generators like AudioGen will likely also make it easier to produce sounds that directly relate to a given image. This is particularly important in regard to multisensory research. Edmiston & Lupyan (2015) showed that stronger perceived congruence between an image and a sound resulted in faster object identity verification judgements. For example, a bird sound speeded visual object identity verification when it was a strong match for the bird presented in the image. AI-generated stimuli likely have room to grow in terms of quality, and will require the same methods for norming that have been applied to images and artwork in the past, but they will undoubtedly expedite the creation of new stimuli that can be used for scientific purposes.

Another other area that has seen a tremendous amount of growth in recent years is the development of methods for studying cognitive processing using virtual and augmented reality (also referred to jointly as extended reality or XR). XR headsets are relatively low in cost for researchers, and allow for tight control of stimuli along with more naturalistic presentations. Rather than the cost, the skillset necessary for programming XR studies tends to be the barrier

for researchers in psychology, given that the software needed is primarily design for game developers rather than scientists. Two game engines are primarily used for developing VR and AR studies, including Unity Game Engine and Unreal Engine. These typically require bespoke code written in C# , C++, or Java, especially for extracting relevant data like participant responses, response times and motion and eye tracking. There are a growing number of online resources available for those interested in creating XR studies using Unity or Unreal, either published (e.g., Brookes et al., 2019; https://github.com/immersivecognition/unity-experiment-framework) or made by scientists and made available on GitHub or OSF, as with the experiment structure created for Chapter 4 (https://github.com/sheaduarte/Multisensory-Causal-Inference-VR). Vizard (by WorldViz) is also a useful python package built specifically for VR experiment creation. Because it was built for scientists, the package has an array of functions that are useful for writing relevant data including response times and eye tracking. Python is also generally a more widely used language by cognitive scientists for building two-dimensional psychophysics experiments and conducting data analyses, making this a more accessible route for many looking to run VR studies. There is a tradeoff between these software options given that Unity and Unreal licenses are free for use in academic research, whereas Vizard is not, though Vizard can be a less labor-intensive option for getting VR experiments off the ground quickly.

Given these technical challenges, Large Language Models (LLMs) are another way that generative AI can support researchers in developing XR experiments. LLMs like OpenAI's ChatGPT, Anthropic's Claude, and Google's Gemini (among many others), can be resources for getting started in programming XR studies, particularly because they can provide answers to programmers at various stages of expertise and can provide customized code. For example, models like these can help a new developer edit code shared online by other researchers to suit

their own needs, and even write new methods and classes for additional functionality. As with all generative AI, the code generated may be imperfect or require manipulation, though it will be a tremendous help for researchers hoping to begin building XR experiments without starting from scratch.

Applications for education and technology

The findings reported in this dissertation underscore the contributions of realistic stimuli and sensory information across modalities to perception, memory, and learning. The findings from Chapters 2 and 3 suggest that presenting information in multiple modalities can improve recognition and associative memory, which are both necessary for learning (Yonelinas 2002, Mitchell & Johnson, 2009). Other learning and memorization techniques aim at increasing attention or facilitating elaboration using mnemonics like a Memory Palace or information mapping (Roediger, 1980). However, across all of our reported experiments, participants were unaware that they were meant to remember objects or form associations between stimuli, suggesting that multisensory encoding can support memory and learning without increasing effort of the part of the observer. While more research will help us further understand why this is the case, these findings do suggest that taking advantage of multiple sensory modalities has immense potential for developing educational methods for improving retention. Educational tools are increasingly utilizing technology including tablets, computers, video games, and even XR, which may already benefit from multimodal presentations (e.g., Reggente et al., 2020). The ways in which multisensory presentations can be integrated into technology for learning purposes and engagement represents and exciting avenue for future research.

Understanding how perception is affected by stimuli presented to multiple modalities is also important to the development of wearable technology that is intended to augment the ways

in which users experience the world around them. It is rare to encounter anyone who does not augment their sensory environment in one way or another using technology: most frequently through the use of noise-attenuating or cancelling earbuds, but also through the use of VR or AR headsets. Novel technologies like these are intended to manipulate the sensory world for users to tune it to be the most useful, interesting, or entertaining that it can be, and integrating the principles of multisensory processing can improve these efforts. For example, when Apple AirPods are used in passthrough mode, environmental sounds are reproduced within the earbuds, though they do so using spatial audio processing that creates the illusion that the sounds are coming from their natural directions, making the experience more immersive and making it easier to locate the visual sources of environmental sounds. Our results from Chapter 4 suggest that integration can be enhanced by exploiting associations that users already have from real-world object-sound associations by using skeuomorphic or true-to-life audiovisual designs in AR experiences that are meant to feel seamless. How object realism affects other cognitive processes like attentional capture would be an interesting avenue of research for future studies and for application in technology.

References

Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., & Lepage, M. (2010). The Bank of Standardized Stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PloS one*, *5*(5), e10773.

Brookes, J., Warburton, M., Alghadier, M., Mon-Williams, M., & Mushtaq, F. (2020). Studying human behavior with virtual reality: The Unity Experiment Framework. *Behavior research methods*, *52*, 455-463.

Bruns, P. (2019). The Ventriloquist Illusion as a Tool to Study Multisensory Processing: An Update. *Frontiers in Integrative Neuroscience*, *13*, 51. https://doi.org/10.3389/fnint.2019.00051

Diana, R. A., Yonelinas, A. P., & Ranganath, C. (2007). Imaging recollection and familiarity in the medial temporal lobe: A three-component model. *Trends in Cognitive Sciences*, *11*(9), 379–386. https://doi.org/10.1016/j.tics.2007.08.001

Edmiston, P., & Lupyan, G. (2015). What makes words special? Words as unmotivated cues. *Cognition*, *143*, 93–100. https://doi.org/10.1016/j.cognition.2015.06.008

Eichenbaum, H., Yonelinas, A. P., & Ranganath, C. (2007). The Medial Temporal Lobe and Recognition Memory. *Annual Review of Neuroscience*, *30*(1), 123–152. https://doi.org/10.1146/annurev.neuro.30.051606.094328

Heikkilä, J., Alho, K., Hyvönen, H., & Tiippana, K. (2015). Audiovisual Semantic Congruency During Encoding Enhances Memory Perfor- mance. *Experimental Psychology*, *62*(2), 123– 130. https://doi.org/10.1027/1618-3169/ a000279

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal Inference in Multisensory Perception. *PLoS ONE, 2*(9), e943. https://doi.org/10.1371/journal.pone.0000943

Lehmann, S., & Murray, M. M. (2005). The role of multisensory memories in unisensory object discrimination. *Cognitive Brain Research*, *24*(2), 326–334. https://doi.org/10.1016/j.cogbrainres.2005.02.005

Macaluso, E., Noppeney, U., Talsma, D., Vercillo, T., Hartcher-O'Brien, J., & Adam, R. (2016). The Curious Incident of Attention in Multisensory Integration: Bottom-up vs. Top-down. *Multisensory Research*, *29*(6–7), 557–583. https://doi.org/10.1163/22134808-00002528

Matusz, P. J., Wallace, M. T., & Murray, M. M. (2017). A multisensory perspective on object memory. *Neuropsychologia*, *105*, 243–252. https://doi.org/10.1016/j.neuropsychologia.2017.04.008

Mitchell, K. J., & Johnson, M. K. (2009). Source monitoring 15 years later: What have we learned from fMRI about the neural mechanisms of source memory? *Psychological Bulletin*, *135*(4), 638–677. https://doi.org/10.1037/a0015849

Moran, Z. D., Bachman, P., Pham, P., Hah Cho, S., Cannon, T. D., & Shams, L. (2013). Multisen- sory Encoding Improves Auditory Recognition. *Multisensory Research, 26* (6), 581–592. https : //doi.org/10.1163/22134808-00002436

Reggente, N., Essoe, J. K., Baek, H. Y., & Rissman, J. (2020). The method of loci in virtual reality: explicit binding of objects to spatial contexts enhances subsequent memory recall. *Journal of Cognitive Enhancement, 4*, 12-30.

Roediger H. L. (1980). The effectiveness of four mnemonics in ordering recall. *Journal of Experimental Psychology: Human Learning and Memory*, 6, 558–567. doi:10.1037/0278-7393.6.5.558

Rohe, T., & Noppeney, U. (2015). Sensory reliability shapes perceptual inference via two mechanisms. *Journal of Vision*, *15*(5), 22. https://doi.org/10.1167/15.5.22

Schneider, T., Engel, A., & Debener, S. (2008). Mul- tisensory identification of natural objects in a two-way crossmodal priming paradigm. *Experi- mental Psychology*, *55*, 121–131.

Slutsky, D. A., & Recanzone, G. H. (2001). Temporal and spatial dependency of the ventriloquism effect: *Neuroreport*, *12*(1), 7–10. https://doi.org/10.1097/00001756-200101220-00009

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: Human learning and memory*, *6*(2), 174.

Shams, L., & Seitz, A. R. (2008). Benefits of multisensory learning. *Trends in Cognitive Sciences*, *12*(11), 411–417. https://doi.org/10.1016/j.tics.2008.07.006

Stein, B. E., Stanford, T. R., & Rowland, B. A. (2020). Multisensory Integration and the Society for Neuroscience: Then and Now. *The Journal of Neuroscience*, *40*(1), 3–11. https://doi.org/10.1523/JNEUROSCI.0737-19.2019

Thelen, A., Talsma, D., & Murray, M. M. (2015a). Single-trial multisensory memories affect later auditory and visual object discrimination. *Cog- nition*, *138*, 148–160. https://doi.org/10.1016/ j.cognition.2015.02.003

Wais, P. E., Wixted, J. T., Hopkins, R. O., & Squire, L. R. (2006). The Hippocampus Supports both the Recollection and the Familiarity Compo- nents of Recognition Memory. *Neuron*, *49*(3), 459–466. https://doi.org/10.1016/j.neuron. 2005.12.020

Van Wanrooij, M. M., Bremen, P., & John Van Opstal, A. (2010). Acquired prior knowledge modulates audiovisual integration. *European Journal of Neuroscience*, *31*(10), 1763–1771. https://doi.org/10.1111/j.1460-9568.2010.07198.x

Yonelinas, A. P. (2002). The Nature of Recollection and Familiarity: A Review of 30 Years of Research. *Journal of Memory and Language*, *46*(3), 441–517. https://doi.org/10.1006/jmla.2002.2864

Yonelinas, A. P., Aly, M., Wang, W.-C., & Koen, J. D. (2010). Recollection and familiarity: Examining controversial assumptions and new directions. *Hippocampus*, *20*(11), 1178–1194. https://doi.org/10.1002/hipo.20864