Mind the Mix: Exploring the Cognitive Underpinnings of Multimodal Interaction in Augmented Reality Systems

May Jorella Lazaro[*]
CX Insight Team, MX Division, Samsung Electronics Co. Ltd., South Korea
ella.lazaro@samsung.com

Sungho Kim
Department of Systems Engineering, Republic of Korea Air Force Academy, South Korea
sunghokim1123@gmail.com

ABSTRACT

Exploring the intricate dynamics of Multimodal Interaction (MMI) in Augmented Reality (AR), this study presents a novel conceptual framework, crafted from a review of cognitive theories. Our research delves into how input modalities, output modalities, and their combinations uniquely influence user experiences in AR environments. Recognizing a gap in the existing MMI literature, especially within the AR context, we propose a conceptual framework to understand these complex relationships. Our framework pinpoints three critical factors: the choice of input modality, the verbal processing code of outputs, and the synergistic effects of input-output combinations. These elements are hypothesized to significantly impact user interaction and performance in AR systems. This work-in-progress not only contributes to the theoretical discourse in HCI but also sets the stage for future empirical investigations, aiming to enhance user-centered design in the evolving field of AR technology.

CCS CONCEPTS

• **Human-centered computing** → Human computer interaction (HCI); Interaction paradigms; Mixed / augmented reality; Human computer interaction (HCI); HCI theory, concepts and models.

KEYWORDS

Multimodal Interaction, Augmented Reality, Cognitive Theories, Input and Output Modalities

ACM Reference Format:
May Jorella Lazaro and Sungho Kim. 2024. Mind the Mix: Exploring the Cognitive Underpinnings of Multimodal Interaction in Augmented Reality Systems. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3613905.3650874

1 INTRODUCTION

As we navigate through an era where the boundaries between the digital and physical worlds are increasingly blurred, Augmented Reality (AR) has emerged as a transformative force. This technology not only reshapes our digital experiences but also redefines Human-Computer Interaction (HCI). AR's capability to overlay digital information onto our physical environment positions it uniquely within the spectrum of eXtended Reality (XR). In contrast to Virtual Reality (VR), which creates a wholly digital environment, the intrinsic dependence of AR on the variabilities of the real-world significantly increases the complexity of its interaction design. Central to this evolution is the exploration of interaction modalities, particularly Multimodal Interaction (MMI). While MMI is known to enhance the efficiency and naturalness of human-system interactions, a significant gap remains in understanding the optimal integration of various interaction modalities within AR systems [6]. This gap is critical, as it directly impacts the effectiveness of AR systems in real-world applications.

The positive influence of MMI on user performance and immersive experiences in XR environments is well-documented [10, 13, 14, 17, 18, 26, 32]. However, the efficacy of MMI is not universally consistent. Some studies suggest that MMI may not always outperform unimodal interactions [32, 33] and may even increase workload or the likelihood of error if not properly implemented [7, 19, 20]. These findings underscore the inherent complexity of MMI, yet there is still a notable gap in efforts to understand the underlying factors that influence its effects.

In this paper, we delve into the cognitive aspects of MMI, proposing that a successful multimodal system should align with the user's cognitive capabilities and limitations. We begin by presenting an overview of cognitive theories relevant to MMI, highlighting their implications for designing AR interactions. Subsequently, we introduce a conceptual framework for MMI in AR, which is informed by a comprehensive review of these theories. The proposed framework elucidates the interplay of various input and output modalities during human-system interaction, with a particular focus on the cognitive processes underlying information processing.

Through this study we were able to come up with several hypotheses on how the combination of different interaction modalities might work and what are the factors that we might need to consider during its implementation. While these hypotheses were not empirically tested within the scope of this study, they lay a foundational groundwork for advancing our comprehension of MMI. This work-in-progress sets the stage for future research to test and build upon these initial insights.

Our goal is to offer insights that not only possibly enhance the effectiveness of MMI but also contribute to the broader discourse in HCI, particularly in creating more intuitive and user-friendly AR systems. This investigation is particularly pertinent to the CHI community, extending discussions about MMI within the context of

[*]Corresponding Author

CHI EA '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0331-7/24/05
https://doi.org/10.1145/3613905.3650874

emerging AR technologies. By examining the cognitive underpinnings of MMI, we aim to contribute to a more nuanced understanding of AR interactions, offering practical insights for designers and developers in optimizing next-generation AR experiences.

2 THEORETICAL REVIEW

2.1 Overview of cognitive theories

Humans are naturally inclined to interact with the world through multiple modalities. In natural human interaction, people instinctively use different ways to transmit and perceive information. Literature in biology, experimental psychology, and cognitive science has consistently shown that human information processing basically involves multisensory perception and integration [9]. In fact, the human brain contains multimodal neurons and specialized multisensory convergence regions that support multimodal processing [21]. With this, several studies assert that multimodal interface systems are advantageous and superior since they support natural human information processing. This assertion is mostly based on the number of theories in cognitive psychology that discuss multiple, modality-specific processing resources [3, 4, 25, 34].

During human-computer interaction, the users' information processing can be broadly divided into three stages: input, processing, and output. These three stages are sometimes referred to as stimulus (or perception), cognition, and response (or action) in some studies [31]. However, it is important to note that these terms can be confusing as the terminology used in this research is interaction focused. In this study, input refers to the stage where users actively participate and transmit information to the system (i.e., user input). Hence, the input stage in this study can be translated into the response stage discussed in most cognitive theories. On the other hand, in this study, output refers to the information produced by the system for the users to perceive (i.e., system output). This then refers to the stimulus or perceptual encoding stage discussed in other cognitive theories. The processing stage, however, refers to the same concept as discussed in most information processing-related theories, which basically means the stage where cognition takes place. There are several psychological theories that have influenced contemporary views of multimodal interaction and interface design. The most relevant ones are working memory theory, dual-coding theory, multiple resource theory, cognitive load theory, and sensory-motor modality compatibility theory. In the following subsections, we provide detailed explanations of each theory.

2.1.1 Working memory theory. Baddeley's working memory hypothesis states that various forms of information correspond to distinct cognitive resources [1]. Working memory refers to the type of memory that humans use to temporarily store information needed for processing, which is usually of limited capacity [1]. According to Baddeley's initially proposed model, a human's working memory is made up of three main components: the central executive, visuo-spatial sketchpad, and phonological loop. The central executive acts as the supervisory system, wherein the overall processing and integration of information take place. The visuo-spatial sketchpad is used for short-term storage of visual-spatial information, while the phonological loop is used for short-term storage of auditory-verbal data. However, it must be noted that according to this theory, in cases where visual information is presented verbally (e.g., printed text), the information is stored in the phonological loop as sub-vocal articulated information. Although the visuo-spatial sketchpad and phonological loop are generally controlled by the central executive, these two components operate mostly independently in terms of lower-level modality processing. After several years, Baddeley [2] updated his model after conducting several empirical investigations. He then added the fourth component which is the episodic buffer, which represents the path between the long-term and working memory. As an implication for MMI, this model posits that human performance can be improved when the interaction involves using multiple modalities from different components (visuo-spatial sketchpad and phonological loop).

2.1.2 Dual coding theory. Another relevant theory for MMI is the dual coding theory proposed by Paivio [24]. In his theory, he suggested that human cognition is mainly composed of two processing systems: the visual (sometimes referred to as the non-verbal system), and the verbal system. The visual system is concerned with storing and processing graphical information, whereas the verbal system deals with storing and processing linguistic information. By its conceptual definitions, the two components are somewhat similar to Baddeley's concept of the visuo-spatial sketchpad and the phonological loop. Paivio [25] asserted that although written text and graphical pictures are both presented as visual stimuli, it is coded and processed in different ways. For example, written text is coded more verbally than visually, whereas pictures can be coded both visually and verbally due to subconscious labeling. According to Paivio [25], these assumptions are widely supported by studies in neuropsychology. It was shown that processing different types of information (verbal and non-verbal) activates different areas of the brain, and the presentation of both types leads to more brain stimulation. Moreover, this theory posits that the processing of verbal stimuli prompts a verbal response, and the processing of nonverbal stimuli prompts a nonverbal response.

2.1.3 Multiple Resource Theory. Multiple resource theory (MRT) proposed by Wickens [34] suggests that humans have a limited set of resources available to mentally process information. These resources are regarded as a pool of energy that can be consumed to perform different mental operations, ranging from sensory-level processing to semantic-level processing [5]. In this model, information processing is divided into three stages: perception, cognition, and responding. The information to be processed is further divided into two different perceptual modalities (visual and auditory), two different processing codes (verbal and spatial), and two different response modalities (manual and vocal). In addition, visual information can also be further classified into two categories: focal and ambient. An illustration of the multiple resource model is presented in Figure 1(c). In contrast with the other theories discussed, Wickens [34] distinguished modalities not only in the processing stage but also in the sensory and response stages. It is assumed that at each stage, different cognitive resources are being used. MRT was initially proposed to predict human performance while executing different types of tasks that involve coordination between the user input and system output modes. According to this theory, tasks that require the same set of resources may be difficult to perform in parallel and can lead to performance deterioration. On the contrary,

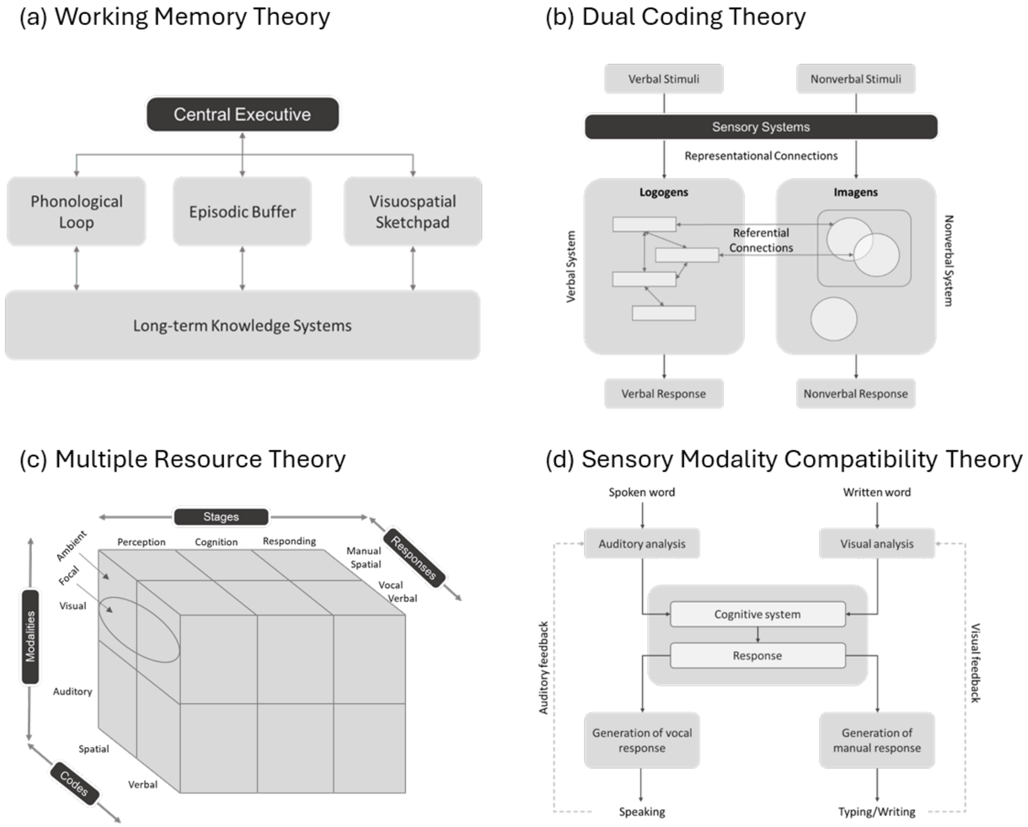

Figure 1: Illustration of theories reviewed. (a) working memory theory diagram adapted from [2], (b) dual coding theory diagram adapted from [25], (c) multiple resource theory model adapted from [33], and (d) sensory-motor modality compatibility theory diagram adapted from [27].

tasks that require accessing different types of cognitive resources (e.g., visual and auditory) can be processed simultaneously with less interference. In summary, MRT proposes that information processing dimensions and components should be taken into account when designing multimodal interfaces, reinforcing coordination and timesharing, and minimizing interference.

2.1.4 Cognitive Load Theory. Cognitive load theory (CLT) is first introduced by Sweller [30]. The main assumption of this theory is that working memory is only able to hold a limited amount of information at a time, therefore in order to maximize learning and performance, the cognitive load (the amount of information that can be stored and processed) must not exceed its limits. When unnecessary demands are imposed on the users, the processing of information becomes overly complex which often leads to loss of information and impaired performance. Cognitive load can be classified into three types: germane, intrinsic, and extraneous load. Germane load refers to the natural demands imposed when constructing schemas and connections that are critical to the learning process. On the other hand, intrinsic load refers to the internal demands brought by the inherent complexity of the task, and extraneous load refers to the non-task-related factors contributing to increasing task complexity (e.g., inappropriate use of modality). According to Sweller [30], there are a lot of factors that influence cognitive load, such as modality and schemata. In addition, several studies have also shown that task demands and complexity, extraneous noise, decision-making, etc. also have a huge impact on cognitive load [12, 16]. Therefore, these factors should also be considered in the investigation of MMI.

2.1.5 Sensory motor modality compatibility theory. The term 'sensory-motor modality compatibility' or sometimes referred to as 'modality compatibility' describes the relationship between input and output modalities and how well they work together [11]. It is believed that when the stimulus and response are mapped under the same modality, it will more likely lead to more compatible and efficient information processing. For example, auditory stimuli paired with a vocal response, or visual stimuli paired with manual response, are regarded as compatible since the motor response leads to the sensory feedback that corresponds to the previously presented stimuli. On the contrary, input-output combinations such as visual stimuli and vocal response, or auditory stimuli and manual response, are regarded as incompatible due to the lack of correspondence between the modalities. This belief is rooted based on the classical ideomotor principle which asserts that actions are selected and initiated based on their expectations of the effects [15]. This means that even

Table 1: Summary of findings from theoretical review

Relevant MMI component	Implications on MMI	Reference theory
Input	Extraneous loads induced by additional decision-making processes (e.g., provision of alternative input choices) can unnecessarily increase task load, which could lead to adverse effects on performance.	Cognitive load theory
Output	The verbal characteristics of the output (verbal or non-verbal) are processed differently, and the presentation of both types of information can lead to higher brain stimulation.	Dual coding theory
Input and Output	The use of different modalities from different processing components (visuo-spatial sketchpad and phonological loop) can lead to improved task performance.	Working memory theory
Input and Output	The use of inputs (manual-spatial, vocal-verbal) and the processing of output modalities (visual and auditory) from different resources can lead to increased efficiency due to parallel processing and time-sharing.	Multiple resource theory
Input-output combination	Information processing would be more efficient when the stimulus and response are mapped under the same modality (e.g., auditory-vocal, visual-manual)	Sensory-motor modality compatibility theory
Input-output combination	Processing of information and response execution could be influenced by the demands and complexity of the task.	Cognitive load theory

though they are processed in different stages, both sensory input and motor output prime each other leading to a combined effect.

2.2 Summary and insights

In summary, based on the cognitive theories reviewed, it can be assumed that efficiency in information processing can depend on the modalities and type of cognitive resources used. More specifically, several factors underlying each specific component of MMI are presumed to influence performance during interactions. The summary of the implications derived based on the assumptions of each cognitive theory on MMI is presented in Table 1.

In terms of input modality, several theories suggest that the type of input modality (whether it is vocal or manual) and the availability of alternative choices (whether it is fixed unimodal or alternate multimodal) can have an impact on the overall task efficiency. Deriving from the assumptions of the cognitive load theory, the increased level of modality choices or freedom during the interaction may actually lead to higher cognitive load and poor performance [28]. From this, it can be deduced that simply incorporating various interaction modalities into a multimodal system does not inherently enhance user performance and experience. Instead, these modalities should be carefully selected and implemented according to appropriate modality mappings.

On the other hand, in terms of output modality, multiple cognitive theories postulate that external information is processed mostly based on its verbal characteristics. This becomes particularly pertinent in AR environments, where digital and physical worlds converge, often resulting in an overlay of multimodal information. Thus, as an implication, it could also be assumed that aside from output modality, the verbal features of the system output could have a significant influence on MMI in AR. Hence, further investigation into the effects of the combined verbal features of multimodal system output may be necessary to confirm these assumptions.

Various cognitive theories also indicate that certain combinations of input-output modalities can lead to better performance, whereas some may lead to interference. Based on multiple resource theory, multimodal systems can lead to inferior performance than unimodal systems if the combinations of input and output modalities and their corresponding dimensions interfere with each other [34]. For example, when presenting identical information using two different modalities (e.g., reading and listening to the same text simultaneously), even though it uses different modality resources, it can lead to decreased performance due to problems with synchronizing information with the same processing code [29]. Therefore, it is crucial to investigate the possible effects of combining different input and output modalities in order to fully understand the implications of MMI.

In most cases, empirical findings support the assumption that multimodal interfaces result in more natural, efficient, reliable, and robust interaction [8, 22, 23]. However, it is important to note that these advantages are not inherent in all multimodal systems. It can be seen through this review that the advantages of MMI are mediated by multiple factors such as the combination of input and output conditions. Apart from this, according to cognitive load theory, the type of task, task complexity, and task setting (single task vs task switching) can also increase cognitive load and influence the performance gain from MMI. It is possible that the combination of some interaction modalities may cause interference due to the nature of some tasks with which concurrent information processing is rather difficult [35]. Thus, investigating MMI components in different task scenarios with varying complexities and demands is also essential.

3 PROPOSED FRAMEWORK

With reference to the general model of the human-computer communication process previously established by Schomaker [28], a

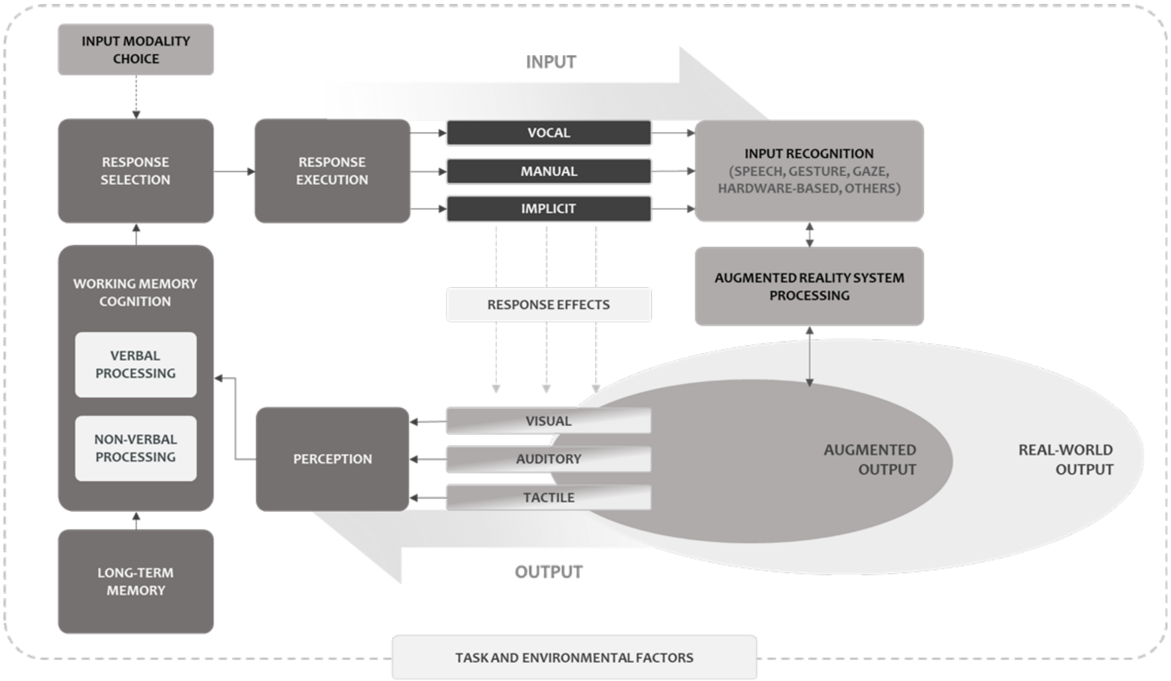

Figure 2: Proposed framework for MMI in AR

conceptual framework for MMI in AR is developed. The original model presents a functional and cognitive parallel between the human and the system. In the proposed model, information is transmitted either through the input flow (user input to system recognition) or the output flow (system output to the user perception). The original framework was further extended by including framework components from the reviewed cognitive theories related to MMI and AR system-related factors.

Human information processing components derived from the original model of human information processing and multiple resource theory [35], such as perception, working memory cognition, long-term memory, response selection, and response execution, were included in the conceptual framework to demonstrate how the users process and produce multimodal information. Furthermore, in alignment with the dual coding theory and the sensory-motor modality compatibility theory: verbal and non-verbal processing. Additionally, the proposed conceptual framework highlights the role of the sensory response effects during multimodal interaction. It is postulated that the feedback produced by the execution of the user input (e.g., vocal input produces an auditory sensory response effect) affects the users’ perception.

AR system-related factors such as user input recognition, AR system processing, augmented output, and input modality choice were integrated into the conceptual framework. Augmented output refers to the stimuli produced by the AR system, which is presented on top of the real-world output. Real-world output encompasses all the stimuli from the real-world environment, whether task-related (e.g., tactile feedback from physical objects) or non-task-related (e.g., environmental noise, ambient lighting). In addition to the primary response types (vocal and manual), implicit user input was included as part of the user input modalities. Implicit user input refers to the involuntary or subconscious responses of the users that are detected and processed by the AR system (e.g., biosignals, facial expressions, and implicit gaze). Input modality choice, on the other hand, refers to the input conditions and restrictions implemented within the AR system (e.g., equivalent multi-input options, assigned multi-input, unimodal input). In this conceptual framework, input modality choice in the AR environment is expected to affect the response selection process of the users.

Lastly, in addition to the main framework components, based on the evidence gathered, the effects of the identified components are heavily influenced by factors related to the nature of the task and environment. It is quite known in the vast literature that task characteristics and environmental factors strongly influence human performance in a myriad of ways. Hence, it is not surprising that the effectiveness and efficiency of MMI can also depend on the number and the type of tasks being performed. Since different tasks and task sets require different forms of interaction, the implementation of different input or output modalities may have distinctive effects depending on the task and environmental factors.

The resulting framework is presented in Figure 2. The diagram illustrates the exchange of information between the user and the AR system using multiple modalities along with the factors that could influence the interaction.

4 DISCUSSION AND CONCLUSION

In this study, we unveil a pioneering conceptual framework for MMI in AR systems, crafted from an in-depth exploration of cognitive theories. Our framework underscores the interplay between various factors in MMI within AR settings, proposing that the combined effects of input and output modalities, along with the context of use, could significantly impact user experience and performance. Moreover, the framework suggests that aside from interaction modalities, specific factors including the choice of input modality, the verbal processing code of outputs, and their integrated mappings—shaped by task requirements and environmental contexts—are also crucial in determining the effectiveness of MMI in AR.

Using the proposed framework, numerous hypotheses can be derived and tested. For instance, to reduce cognitive load during complex tasks, offering a limited selection of input methods (such as hand gestures and voice commands) could foster more effective interactions than providing a wide array of suboptimal options (like hand gestures, head gestures, voice commands, eye gaze, and tangible devices). Another potential hypothesis is that restricting verbal outputs to a single modality (for example, combining visual text with ambient sound) may lead to better user performance than presenting multimodal outputs that share the same processing code (such as combining visual text with verbal sound). This hypothesis suggests that designing AR interactions aligned with the ideal combination of interaction modalities, considering its specific attributes, could optimize and enhance both user performance and experience.

Researchers can utilize this framework to design studies that systematically manipulate these variables—input modalities, output processing codes, and environmental conditions—to observe their collective impact on user performance and satisfaction in AR. By creating scenarios that vary these factors, such as comparing user experiences in quiet museum settings versus bustling city tours, researchers can identify patterns and best practices for MMI design in AR, catering to diverse applications and user needs. Observed effects from various variable combinations and contexts could be synthesized to generate more systematic guidelines that could help AR designers and developers in the future. This holistic approach not only broadens our understanding of MMI's complexities in AR but also guides the development of more adaptive, user-centric AR technologies.

Despite the valuable insights offered, our study acknowledges its limitations. The framework's development was based on a selective review of cognitive theories, which may not cover the full scope of relevant literature, potentially resulting in gaps in our insights. Notably, one significant oversight is the omission of spatial cognition—a crucial element in AR. Integrating insights from theories on spatial perception could greatly enrich our framework, as understanding spatial relationships is fundamental to AR environment design and significantly affects user cognition during interaction. Furthermore, the qualitative nature of our analysis, along with the reliance on a theoretical foundation for our framework, could introduce biases and narrow the range of our conclusions.

Nevertheless, this work-in-progress lays a critical foundation for understanding the cognitive underpinnings of MMI in AR systems. Our study not only enriches the discourse in HCI but also provides a structured approach to examining multimodal interactions in AR. Moving forward, we aim to expand our framework through comprehensive reviews of additional cognitive theories and empirical research. Furthermore, we intend to empirically test and refine the hypotheses generated from our framework, thereby contributing to the optimization of MMI in AR systems. We are excited to further explore these hypotheses through rigorous testing, aiming to refine and validate our framework.

In conclusion, this study marks a significant step towards unraveling the complexities of MMI in AR. By bridging cognitive theory with practical HCI design, we offer a nuanced perspective that promises to guide future research and innovation in AR technology. As we continue to explore these dynamics, our work aims to foster more intuitive, effective, and user-friendly AR systems, pushing the boundaries of how we interact with technology and shaping the future of AR experiences.

REFERENCES

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

[2] A. Baddeley. 2000. The episodic buffer: a new component of working memory? Trends in cognitive sciences 4, 11 (2000), 417-423.

[3] A. Baddeley. 2003. Working memory: looking back and looking forward. Nature reviews neuroscience 4, 10 (2003), 829-839.

[4] A. D. Baddeley and G. Hitch. 1974. Working memory. In Psychology of learning and motivation, Vol. 8. Academic Press, 47-89.

[5] M. D. Basil. 2012. Multiple resource theory. In Encyclopedia of the Sciences of Learning. 2384-2385.

[6] M. Billinghurst, H. Kato, and S. Myojin. 2009. Advanced interaction techniques for augmented reality applications. In Proc. International Conference on Virtual and Mixed Reality. Springer, Berlin, Heidelberg, 13-22.

[7] S. A. Brewster, P. C. Wright, and A. D. Edwards. 1994. The design and evaluation of an auditory-enhanced scrollbar. In Proc. SIGCHI Conference on Human Factors in Computing Systems. ACM, 173-179.

[8] J. L. Burke *et al.* 2006. Comparing the effects of visual-auditory and visual-tactile feedback on user performance: a meta-analysis. In Proc. 8th international conference on Multimodal interfaces. ACM, 108-117.

[9] G. Calvert, C. Spence, and B. E. Stein (Eds.). 2004. The handbook of multisensory processes. MIT Press.

[10] P. R. Cohen *et al.* 1997. Quickset: Multimodal interaction for distributed applications. In Proc. fifth ACM international conference on Multimedia. ACM, 31-40.

[11] E. Fintor, D. N. Stephan, and I. Koch. 2018. Emerging features of modality mappings in task switching: Modality compatibility requires variability at the level of both stimulus and response modality. Psychological Research 82, 1 (2018), 121-133.

[12] N. Hollender *et al.* 2010. Integrating cognitive load theory and concepts of human–computer interaction. Computers in human behavior 26, 6 (2010), 1278-1288.

[13] W. Hürst and K. Vriens. 2016. Multimodal feedback for finger-based interaction in mobile augmented reality. In Proc. 18th ACM International Conference on Multimodal Interaction. ACM, 302-306.

[14] S. Irawati *et al.* 2006. An evaluation of an augmented reality multimodal interface using speech and paddle gestures. In International Conference on Artificial Reality and Telexistence. Springer, Berlin, Heidelberg, 272-283.

[15] W. James. 1890. The Principles of Psychology. Henry Holt and Company, New York.

[16] P. A. Kirschner, P. Ayres, and P. Chandler. 2011. Contemporary cognitive load theory research: The good, the bad and the ugly. Computers in Human Behavior 27, 1 (2011), 99-105.

[17] M. J. Lazaro *et al.* 2021. Interaction Modalities for Notification Signals in Augmented Reality. In Proc. 2021 International Conference on Multimodal Interaction. ACM, 470-477.

[18] M. Lee *et al.* 2013. A usability study of multimodal input in an augmented reality environment. Virtual Reality 17, 4 (2013), 293-305.

[19] A. B. Naumann and I. Wechsung. 2008. Developing usability methods for multimodal systems: The use of subjective and objective measures. In Proc. International Workshop on Meaningful Measures: Valid Useful User Experience Measurement (VUUM), 8-12.

[20] A. B. Naumann, I. Wechsung, and J. Hurtienne. 2010. Multimodal interaction: A suitable strategy for including older users? Interacting with Computers 22, 6 (2010), 465-474.

[21] S. Oviatt and P. Cohen. 2022. The Paradigm Shift to Multimodality in Contemporary Computer Interfaces. Springer, Cham, Switzerland.

[22] S. Oviatt. 1996. Multimodal interfaces for dynamic interactive maps. In Proc. SIGCHI conference on Human factors in computing systems. ACM, 95-102.

[23] S. Oviatt *et al.* 2000. Designing the user interface for multimodal speech and pen-based gesture applications: State-of-the-art systems and future research directions. Human-computer interaction 15, 4 (2000), 263-322.

[24] A. Paivio. 1986. Dual coding and episodic memory: Subjective and objective sources of memory trace components. In Memory and cognitive capabilities:

Symposium in memoriam of Hermann Ebbinghaus. North Holland, Amsterdam, 225-236.

[25] A. Paivio. 1991. Dual coding theory: Retrospect and current status. Canadian Journal of Psychology/Revue canadienne de psychologie 45, 3 (1991), 255.

[26] I. Rakkolainen *et al.* 2021. Technologies for Multimodal Interaction in Extended Reality—A Scoping Review. Multimodal Technologies and Interaction 5, 12 (2021), 81.

[27] S. Schaeffner, I. Koch, and A. M. Philipp. 2016. The role of sensory-motor modality compatibility in language processing. Psychological Research 80, 2 (2016), 212-223.

[28] L. Schomaker. 1995. A taxonomy of multimodal interaction in the human information processing system.

[29] W. Schnotz, M. Bannert, and T. Seufert. 2002. Toward an integrative view of text and picture comprehension: visualization effects on the construction of mental models. In The psychology of science text comprehension, 385-416.

[30] J. Sweller. 1988. Cognitive load during problem solving: Effects on learning. Cognitive science 12, 2 (1988), 257-285.

[31] K. M. Stanney *et al.* 2021. eXtended reality (XR) environments. In Handbook of human factors and ergonomics. 782-815.

[32] E. Triantafyllidis, C. McGreavy, J. Gu, and Z. Li. 2020. Study of multimodal interfaces and the improvements on teleoperation. IEEE Access 8 (2020), 78213-78227.

[33] A. Varghese. 2020. Exploring Bi-modal Feedback in Augmented Reality. In IndiaHCI’20: Proceedings of the 11th Indian Conference on Human-Computer Interaction. ACM, 55-61.

[34] C. D. Wickens. 2002. Multiple resources and performance prediction. Theoretical issues in ergonomics science 3, 2 (2002), 159-177.

[35] C. D. Wickens and C. M. Carswell. 2021. Information processing. In Handbook of human factors and ergonomics, 114-158.