Exploring Diminishing Returns for Multimodal Dynamic Event Expression in Mixed Reality

Aochen Jiao[a], Yuxuan Li[b], Youpeng Zhang[c], Hsin-Yi Wang[d,1], Jiaxuan Wang[e], Weitao Xu[d,*], Eduardo Velloso[a], Nuwan Janaka[d], Shengdong Zhao[d,*]

[a] *The University of Sydney, Sydney, NSW, Australia*
[b] *Wuhan University of Technology, Wuhan, Hubei, China*
[c] *The Hong Kong Polytechnic University, Hong Kong, China*
[d] *City University of Hong Kong, Hong Kong, China*
[e] *University of Southern California, Los Angeles, CA, USA*

Abstract

Expressing dynamic events in Mixed Reality (MR) requires users to communicate continuous spatiotemporal information, including entity trajectories, interactions, and temporal dynamics. This task is fundamentally constrained by the linearization bottleneck: human mental models preserve rich relational structures holistically, yet conventional output channels, such as speech, force them to be serialized into sequential streams. Multimodal interaction offers a promising solution by distributing information across complementary channels, but it remains unclear how different multimodal strategies perform as spatiotemporal complexity increases. This paper addresses these two knowledge gaps: the lack of a systematic analysis of decomposing dynamic events and the absence of empirical evidence comparing multimodal expression strategies under controlled-complexity conditions. We propose a structured analysis that decomposes dynamic events into fundamental infor-

*Corresponding authors: Shengdong Zhao (shengdong.zhao@cityu.edu.hk), Weitao Xu (weitaoxu@cityu.edu.hk).

Email addresses: aochen.jiao@sydney.edu.au (Aochen Jiao), simon2016@whut.edu.cn (Yuxuan Li), 26130336r@connect.polyu.hk (Youpeng Zhang), hsinyi.wang@my.cityu.edu.hk (Hsin-Yi Wang), wangjiax@usc.edu (Jiaxuan Wang), weitaoxu@cityu.edu.hk (Weitao Xu), eduardo.velloso@sydney.edu.au (Eduardo Velloso), nuwanj@u.nus.edu (Nuwan Janaka), shengdong.zhao@cityu.edu.hk (Shengdong Zhao)

[1]Deceased.

mation blocks and establishes a complexity taxonomy. Building on this analysis, we design sequential and parallel multimodal expression strategies and evaluate them alongside sequential (unimodal) speech as complete interaction strategies in a controlled user study using abstract geometric micro-worlds. Under error-free speech–gaze fusion provided through a Wizard-of-Oz setup, parallel multimodal input maintains a substantial completion-time advantage across the tested event levels. However, its subjective benefits decrease as coordination demands rise. When events include asynchronous timing or causal interactions, parallel input is no longer rated significantly higher than sequential multimodal input, though this does not imply equivalence. These findings suggest a boundary condition for the subjective benefits of the parallel strategy, providing design implications for adaptive multimodal MR systems.

Keywords: dynamic events, expression, multimodal interaction, mixed reality, framework

---

1. Introduction

Mixed Reality (MR) has long been envisioned as more than a display technology. It is increasingly becoming a dynamic medium for thinking, communicating, and collaborating. As Bret Victor and others have argued (Victor, 2014), dynamic and interactive representations can fundamentally change how people reason about complex phenomena, from a basketball coach coordinating player movements to a scientist explaining orbital mechanics or a teacher illustrating cause-and-effect relationships. As MR matures, this vision is becoming increasingly practical, further reinforced by recent advances in LLM-supported spatial computing that make dynamic spatial communication richer and more actionable (Du et al., 2026). This trajectory aligns with the vision of *heads-up computing*, which argues for matching input and output channels to human capabilities rather than to device form factors, so that computing support stays embedded in natural, heads-up activity (Zhao et al., 2023). Applications such as asynchronous remote instruction (Piumsomboon et al., 2019; Wang et al., 2020a) and in-situ animation authoring (Suzuki et al., 2020; Wang et al., 2021) increasingly require users not merely to point at static objects, but to generate, describe, and reconstruct complex dynamic events.

Yet a fundamental bottleneck stands in the way. Humans perceive dynamic events as holistic, concurrent spatiotemporal flows (Zacks and Tversky, 2001; Johnson-Laird, 1983). When imagining two cars colliding, for instance, we simultaneously visualize trajectories, speeds, and the moment of impact – a rich, parallel mental representation. The trouble arises when users try to externalize this representation to a computational system. Traditional channels, whether speech or text, are inherently linear and discrete (Levelt, 1993), forcing users to serialize high-dimensional spatial information into a low-dimensional sequential stream. This "linearization problem" causes information loss and imposes a high mental workload, especially when describing complex, multi-object interactions (Levelt, 1981).

Multimodal interaction offers a compelling solution: because speech, gesture, and gaze draw on largely separate cognitive resources, users can in principle transmit information across these channels simultaneously, expanding expressive capacity beyond what any single modality permits (Oviatt and Cohen, 2000). Early seminal works such as Bolt's "Put-That-There" (Bolt, 1980) demonstrated the power of such synergistic combinations for discrete spatial commands. However, dynamic event expression is not a momentary command but a sustained, real-time performance. For this reason, the question is not whether multimodal input is beneficial in principle, but how to harness it effectively for sustained spatiotemporal communication. An intuitive extension is *parallel multimodal* expression, for instance, narrating an event while simultaneously gesturing the trajectories involved, which distributes complementary information across concurrent channels to maximize expressive capacity. Yet this also introduces executive overhead: users must continuously decide what information goes to which modality, keep speech and gesture temporally aligned, and maintain narrative coherence all at once. This overhead, which we term *synchronization cost*, may escalate as spatiotemporal complexity grows, eventually offsetting the capacity advantage of parallelism. An alternative is *sequential multimodal* expression, which retains speech as the dominant modality while scaffolding the expression process to reduce coordination demands so that complementary modalities are selectively invoked for designated information to exploit their representational strengths without demanding continuous real-time synchronization.

These two strategies capture plausible ways of organizing multimodal expression. However, applying it to the expression of dynamic events reveals two main knowledge gaps. First, existing multimodal interaction models are largely designed for discrete commands and lack a more systematic analysis

for decomposing and analyzing continuous dynamic spatiotemporal events. Without such an analysis, investigations into multimodal expression strategies remain ad hoc, with no principled basis for structuring comparisons or interpreting results. Second, while the preceding discussion motivates two distinct multimodal expression strategies, we lack empirical evidence that disentangles their trade-offs under controlled complexity conditions. We do not yet know when or why the effort of synchronization overwhelms the theoretical benefits of parallelism.

If MR is to realize its potential as a dynamic medium, where people can fluidly describe, discuss, and manipulate complex events, we need a principled, empirically grounded understanding of how humans externalize dynamic mental models. In this work, we attempt to address the above knowledge gaps by answering the following research questions:

- **RQ1:** How can continuous dynamic events be conceptualized to systematically investigate expression strategies across varying spatiotemporal complexity in MR?
- **RQ2:** How do *sequential speech* (i.e., unimodal speech), *sequential multimodal*, and *parallel multimodal* strategies compare in expressing dynamic events in MR, and how do their relative benefits and trade-offs in efficiency, mental workload, and user experience change as spatiotemporal complexity increases?

To answer these questions, we adopt Heider and Simmel's abstract microworld paradigm (Heider and Simmel, 1944), using geometric primitives to simulate dynamic events and isolate spatiotemporal complexity from domain-specific semantics. We first establish our structured analysis, which decomposes dynamic events to fundamental information blocks and defines an events taxonomy of spatiotemporal complexity (Section 3). We then design *sequential multimodal* and *parallel multimodal* strategies (Section 4) and evaluate them in a controlled user study, revealing diminishing subjective benefits as events require more spatiotemporal coordination (Section 5). Finally, we interpret these findings through a cognitive theoretical lens to derive design implications for adaptive multimodal MR systems (Section 6).

This work attempts to take one step in that direction. The contributions of this paper are threefold: First, we present a structured analysis that decomposes continuous spatiotemporal flows into fundamental information

units and establishes a complexity taxonomy, providing the theoretical foundation needed to systematically study multimodal expression in MR. Second, by comparing three complete interaction strategies, we show empirically that, in the tested MR setting and with error-free speech–gaze fusion, parallel input keeps its completion-time advantage while its subjective advantages over the sequential multimodal strategy fade as events require more coordination. Once events involve asynchronous timing or causal interaction, we find no significant difference between the two strategies in subjective ratings. Third, we offer theoretical insights into cognitive constraints that may arise when coordinating concurrent multimodal expression, identifying how increasing coordination demands may limit the subjective benefits of parallel input and deriving design implications for adaptive scaffolding in future context-aware MR systems.

2. Related Work

Our research is situated at the intersection of multimodal interaction in MR, the linearization constraints of language production, and cognitive theories of information processing.

2.1. Multimodal Interaction in MR

Multimodal interaction integrates complementary channels such as speech, gesture, and gaze to support more natural and expressive interaction (Oviatt and Cohen, 2000; Jaimes and Sebe, 2007). Wickens' Multiple Resource Theory (MRT) explains this benefit: distributing demands across separable resource pools can reduce cross-modal interference (Wickens, 2002). Prior MR systems apply this principle through three common pairings: (1) gaze and gesture separate perceptual acquisition from motor execution, with gaze locating targets and the hands manipulating them (Piumsomboon et al., 2019; Pfeuffer et al., 2017; Lystbæk et al., 2024; Turkmen et al., 2024); (2) speech and gesture separate verbal from spatial representation, with speech conveying semantic commands and gesture specifying spatial information (Bolt, 1980; Irawati et al., 2006; Tse et al., 2007; Kaiser et al., 2003; Weimer and Ganapathy, 1989); and (3) gaze and speech combine perceptual grounding with verbal semantics, with gaze resolving spatial ambiguity and speech conveying categorical and relational information (Khan et al., 2026; Wang et al., 2024; Lee et al., 2024). Despite these different pairings, these systems primarily support *discrete, atomic* operations—momentary selections, rigid-body

manipulations, or single-shot commands—whose coordination is confined to a brief, bounded time window.

A broader paradigm that treats multimodal input as a first-class requirement is *heads-up computing* (Zhao et al., 2023), which proposes matching input and output channels to human capabilities: a head-piece for visual and audio output, and natural input through voice, subtle finger micro-gestures, gaze, and head motion. This paradigm has since been adopted by multimodal MR systems that combine voice with gaze or subtle manual input to keep users engaged with the physical world (Cai et al., 2025; Lee et al., 2024). Our work inherits this input-side orientation, but pushes it from discrete, momentary commands toward sustained expression of dynamic events, where the coordination of voice, gaze, and gesture over time—rather than the availability of each channel—becomes the central challenge.

Recent advances in AI-assisted XR authoring make the complex, dynamic scenarios considered here increasingly plausible. Systems now combine natural language and direct manipulation with generative models or LLMs to create and modify interactive worlds (Rosenberg et al., 2024; Zhang et al., 2024; De La Torre et al., 2024; Lee et al., 2025; Chen et al., 2025; Tütüncü et al., 2026). These systems chiefly address accessible, reliable content generation, but offer limited discussion of the user-side cognitive demands involved in coordinating different modalities during expression of complex information.

Accordingly, expressing dynamic events is a different cognitive endeavor from issuing discrete commands. While the technical feasibility of combining multiple input modalities in MR is well established (Bashar et al., 2026), the user's cognitive capacity to orchestrate such combinations—maintaining narrative logic, causal structure, and temporal coherence simultaneously—during complex spatiotemporal expression remains largely underexplored.

2.2. Linearized Expression of Dynamic Events

Expressing dynamic events requires translating internal mental representations into external output. Johnson-Laird's theory of mental models (Johnson-Laird, 1983) suggests that people simulate dynamic events as holistic flows, while Hierarchical Event Segmentation Theory (EST) (Zacks and Tversky, 2001; Kurby and Zacks, 2008) explains how they parse these flows into part-whole relations, from fine-grained sensory changes (e.g., object speed) to coarse-grained goals (e.g., completing a transfer task).

Levelt's model of speech production (Levelt, 1993) describes the conversion of these hierarchical, high-dimensional thoughts into speech as *lineariza-*

tion: serializing concurrent information into a temporal sequence of words. This creates two bottlenecks when describing continuous motion. First, categorical language segments fluid trajectories into discrete paths (e.g., describing a spiral as "round, then down, then round again"), requiring users to repeatedly sample and verbalize analog motion (Talmy, 2000). Second, speech loses analog precision: although effective for logic, it conveys metric changes in speed, curvature, and acceleration inefficiently (McNeill, 1992), requiring lengthy descriptions such as "it starts slow, then speeds up a bit, then curves sharply." The resulting workload encourages the use of alternative output channels. Gesture- and sketch-based tools already use such channels for continuous motion: users specify how objects move by sketching or demonstrating the motion directly, instead of describing it (Davis et al., 2008; Santosa et al., 2013; Suzuki et al., 2020; Wang et al., 2021). These tools, however, mainly treat gesture or sketching as the only input and say little about how it is coordinated with concurrent speech.

2.3. Multimodal Expression and Cognitive Trade-offs

To address the limitations of linearization, *parallel multimodal* expression emerges as a theoretical solution. McNeill's Growth Point Theory (McNeill, 1992) posits that speech and gesture co-originate from a single idea unit and are inherently designed for synchronous production (Kita, 2000), while Wickens' MRT (Wickens, 2002, 2008) provides the resource-level rationale: distributing demands across verbal/vocal and spatial/manual pools minimizes cross-modal interference. Bodily actions can further serve epistemic functions (Kirsh and Maglio, 1994), enabling cognitive offloading (Risko and Gilbert, 2016) by physically tracing trajectories rather than laboriously verbalizing them. However, the shift to parallel production may incur *synchronization cost*: even when individual modalities draw on separable perceptual-motor resources, their concurrent initiation competes for the central procedural resource that serializes action planning across modalities (Salvucci and Taatgen, 2008). Under complex spatiotemporal demands, this cost may escalate to outweigh the cognitive benefits of resource distribution. Parallel expression also has a physical cost. Continuous mid-air gestures tire the user's arms, an effect often called "gorilla arm syndrome"; and this fatigue increases as users hold their arms higher and for longer durations (Hincapié-Ramos et al., 2014; Jang et al., 2017; Li et al., 2024).

Heads-up computing responds to this constraint by favouring voice combined with subtle finger micro-gestures over sustained mid-air movement,

a preference that is itself instructive: it suggests that continuous parallel expression may not be physically sustainable even when it is cognitively tractable. This observation reinforces our motivation for treating the sequential and parallel strategies as competing options rather than assuming that more simultaneity is always better.

An alternative is a *sequential multimodal* expression, in which complementary modalities are invoked in structured alternation rather than simultaneously. This strategy has a well-established counterpart in the system-level design space proposed by Nigay and Coutaz (Nigay and Coutaz, 1993). Speech remains dominant to preserve narrative continuity and causality, while selectively invoking other modalities for designated information to exploit their representational strengths. Oviatt's studies (Oviatt, 1999; Oviatt and Cohen, 2000) found that users naturally gravitate toward sequential production as task complexity increases, and external cognition research (Scaife and Rogers, 1996; Kirsh, 2010) suggests that structuring interaction into discrete steps scaffolds the process—each step externalizes a portion of the mental model, converting concurrent coordination into sequential planning.

These two strategies represent competing solutions to the representational mismatch. Parallelism maximizes the amount of information conveyed at once via separable resource pools but may incur a synchronization cost that grows with coordination demands. In contrast, sequential alternation reduces coordination overhead. It converts concurrent spatiotemporal relations into step-by-step actions, though this may increase total expression time. The open question is at what level of coordination demand the potential costs of parallel expression outweigh its expressive advantage. If such costs mainly take the form of sustained cognitive strain, their impact may be more visible in subjective experience than in objective completion time, which we explore further in our study.

In short, prior work shows that combining modalities improves discrete commands. It also provides theoretical reasons to expect both benefits and costs during sustained expression. However, we do not know how these two strategies compare in efficiency and user experience as events require greater coordination. Our study attempts to address this gap directly. We compare *sequential speech*, *sequential multimodal*, and *parallel multimodal* expression across four controlled event types.

3. Methodology: Task Paradigm and Event Modeling

To overcome limitations in prior work, specifically the lack of structured analysis for spatiotemporal complexity in dynamic events, we developed an analytical approach that decomposes dynamic events into elemental information blocks and guides the construction of controlled event hierarchies.

Scientific investigation of complex phenomena often demands abstraction to isolate and manipulate the basic variables of interest. In dynamic event expression, abstraction serves two purposes: identifying the elemental cognitive units users must externalize, and enabling precise experimental control over event complexity without the confounding influence of real-world semantics. To meet these requirements, an effective abstraction must be both cognitively grounded, reflecting how humans perceive and represent events, and operationally tractable, meaning its elements can be expressed through observable human output channels.

The Heider and Simmel abstract micro-world paradigm (Heider and Simmel, 1944) satisfies both requirements. This reductionist approach reconstructs complex real-world dynamics as precisely defined events governed by specific spatiotemporal and causal relationships, while effectively isolating these factors from domain-specific semantic interference.

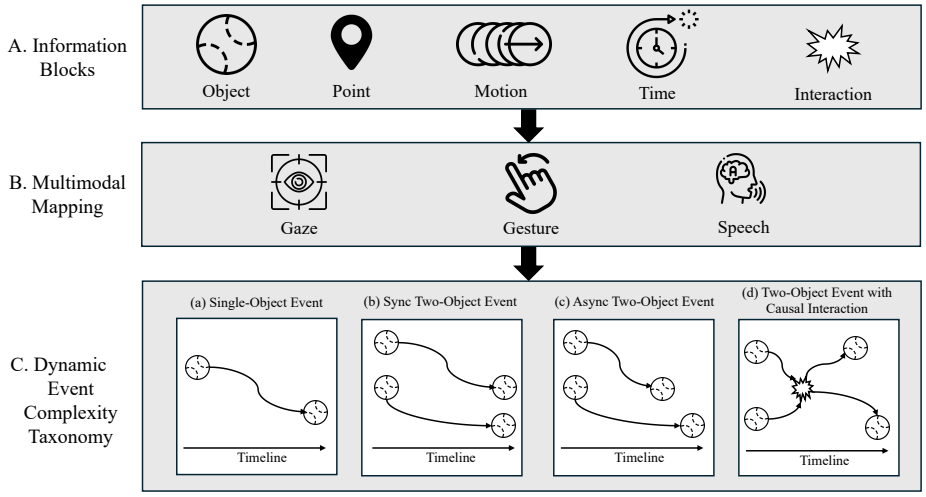


Figure 1: The proposed structured analysis for multimodal expression of dynamic events. (A) Information Blocks: The elementary cognitive units of a dynamic event. (B) Multimodal Mapping: A theoretical map distributing these blocks to human output channels based on cognitive affordances. (C) Dynamic Event Complexity Taxonomy: Four levels of dynamic events constructed by synthesizing these multimodal streams.

3.1. Decomposition: Elemental Information Blocks

When describing a dynamic event unfolding in 3D space, what must a user actually communicate? At minimum, they need to specify what entities are involved, where they are located, when and how they move, and whether they affect one another. These practical demands naturally give rise to five elemental information blocks: object, point, time, motion, and interaction. This decomposition shares a structural resemblance with the Event Indexing Model (EIM) (Zwaan et al., 1995), a well-established cognitive framework that identifies the core dimensions along which humans perceive and mentally represent dynamic events: protagonist, space, time, causality, and intentionality. However, our decomposition diverges from the EIM in an important respect. The EIM was developed to explain narrative comprehension and places particular emphasis on intentionality – a dimension that reflects uniquely human interpretations of agency and goal-directedness. Our focus, by contrast, is on the objective, observable properties of events: what physically happens, where, and when. We do not attempt to model internal states or infer intent.

Similar decompositions also appear in applied AR authoring systems, which represent events and instructions through closely related object, location, time, motion/action, and interaction/logic information (Wang et al., 2020b; Shi et al., 2025), providing practical support for our information block abstraction. We now define each block in terms of what a user must produce to communicate a dynamic event.

3.1.1. Object

An *object* is the entity that acts or is acted upon in an event – the "what" of the scene. It carries semantic identity (e.g., "a red ball") and state attributes. In an MR environment, an object must be explicitly instantiated before it can be referenced by other blocks.

3.1.2. Point

A *point* defines a specific 3D locus within the scene – the "where." It can represent either an absolute position in world coordinates (e.g., a fixed reference location) or a relative offset anchored to another entity (e.g., "behind the object"), accommodating flexible spatial referencing.

3.1.3. Time

Time situates an event along the temporal axis – the "when." This block discretizes continuous time into actionable units: specific timestamps (start and end points) for synchronization, or defined intervals that specify the duration of an action.

3.1.4. Motion

Motion captures how an entity moves through space over time – the "how." We define it as a continuous spatiotemporal trajectory bounded by a specific duration, which converts an abstract goal into a traceable, enactable movement.

3.1.5. Interaction

An *interaction* defines a discrete logical node at which one object influences another – for instance, a collision or a trigger signal. Unlike motion, which captures continuous spatiotemporal evolution, this block represents a distinct causal event that binds independent object trajectories into a coherent relational structure.

3.2. Modality Analysis and Information Mapping

Having decomposed dynamic events into elemental information blocks, we propose a mapping strategy that distributes these units across gaze, gesture, and speech. These three modalities are a natural choice for multimodal MR interaction: they are the primary channels through which users express intent in natural contexts, and they are broadly complementary – each offering affordances the others lack (Wickens, 2008). The mapping is guided by the specific cognitive properties of each channel.

3.2.1. Speech

Speech is uniquely suited for encoding categorical and abstract concepts (Pinker, 2003), and plays a dual role in our analysis framework. As the primary channel for logic, it carries object identity (e.g., “red ball”) and interaction logic (e.g., “collision”), which require the semantic precision that spatiotemporal modalities cannot provide. Speech also serves as a universal fallback: although its linear, sequential nature makes it inefficient for describing continuous spatial trajectories (Levelt, 1993), it is the most expressive channel overall. When gesture or gaze functions imprecisely or fail, users can revert to verbal description to convey motion, temporal information, or spatial points, though at the cost of a higher mental workload.

3.2.2. Gesture

Gestures provide a natural way to mimic physical actions. In our analysis, we focus on deictic, action, spatial, and command gestures (McNeill, 1992) because they are broadly applicable across cultures and languages. Within our framework, gestures serve two roles. First, they support enactment: users express motion paths and spatial points through hand movements. A single gesture can convey multiple layers of information at once, since the traced path communicates spatial form while the speed and rhythm of the movement convey temporal information and other attributes. Second, gestures support direct manipulation. Once abstract information blocks, such as time, are visualized, for example, as tracks on a timeline, users can physically grasp and adjust them. This allows gestures to bridge the gap between ephemeral performance and precise editing.

3.2.3. Gaze

Gaze offers a rich set of interaction features, including pursuit, fixation, and saccadic patterns, each with distinct implications for system design. In this analysis, we focus specifically on fixation because it most directly captures where the user is attending at any given moment and avoids well-known issues such as the Midas touch problem and fatigue associated with using gaze transitions as intentional control signals (Duchowski, 2017). Fixation serves as a stable anchor for spatial coordinates, offloading simple pointing tasks from the hands. Specifically, it naturally aligns with the focus of attention during verbal description, enabling users to transmit point information through fixation while simultaneously conveying semantic details through speech. This supports efficient parallel processing across channels.

3.3. Dynamic Event Complexity: A Structured Analysis

To systematically investigate how increasing event complexity affects multimodal expression, we identify four basic event types that span a range of cognitive demands. This structured analysis is not exhaustive. Rather, it focuses on some of the most basic dynamic events that arise in 3D interactive environments. To avoid confounding our complexity manipulation, we limit our scope to two objects, as tracking additional entities would add unwanted variables. Because these event types are elemental by design, we believe they provide a principled foundation for analyzing and discussing richer, more complex scenarios. The four types follow an ordered progression, though they do not form a single linear scale. Each type introduces a

distinct coordination demand: a second object, asynchronous timing, or a causal relation. Throughout this paper, *complexity* refers to this combined increase in coordination demands.

3.3.1. Single-Object Event

This event involves a single entity executing a continuous spatiotemporal trajectory (Fig. 1(a)). Structurally, it is defined by one *object* block linked to one *motion* block. This level isolates the baseline demand on the spatial/manual response pool required to generate continuous spatiotemporal information for a single entity, devoid of any output competition.

3.3.2. Synchronous Two-Object Event

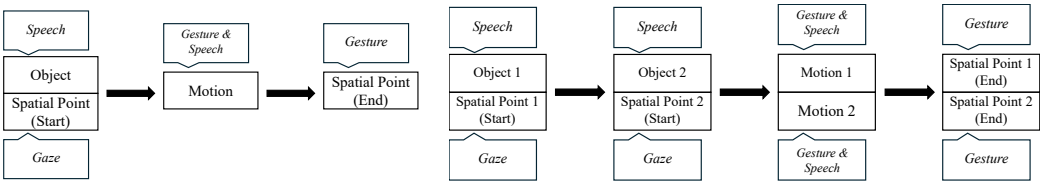
This event introduces concurrent demands on the spatial/manual response pool, occurring when two distinct *objects* execute independent *motions* concurrently within an identical temporal span (Fig. 1(b)). Structurally, it demands generating two *object* and *motion* blocks simultaneously. While this requires maintaining two distinct output streams, the strict temporal synchronization allows efficient parallel allocation of the response pool, thereby reducing interference at the response stage compared to independently controlled actions.

3.3.3. Asynchronous Two-Object Event

This event involves two *objects* moving in non-overlapping or staggered time intervals (Fig. 1(c)). Unlike synchronous motion, the lack of temporal alignment prevents efficient parallel allocation of the spatial/manual response pool. Instead, this imposes additional demands on the central cognitive stage for scheduling and coordination between independent streams.

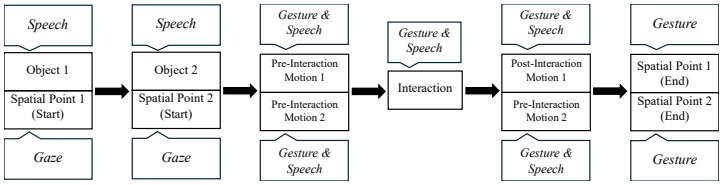
3.3.4. Two-Object Event with Interaction

This event links the spatiotemporal states of two objects via an *interaction* block (Fig. 1(d)). This elevates the demand from processing independent properties to processing relations at the central-processing stage. Defining a causal dependency requires integrating spatial precision, timing, and logical rules simultaneously, imposing a significantly higher load on central processing resources than executing unconnected motions.



(a) Flow for single-object events.

(b) Flow for two-object events.



(c) Flow for two-object events with causal interaction.

Figure 2: Expression flows for *Parallel Multimodal*. Arrows indicate temporal progression between information blocks, and modality callouts identify the channel or channels used for each block. Panels show (a) single-object, (b) two-object, and (c) two-object events with causal interaction; modalities named within the same step are produced concurrently.

4. Designing Multimodal Expression Strategies

We designed and implemented two multimodal expression strategies in an MR prototype. These designs were informed by our analysis of dynamic events and modality affordances (Section 3), as well as feedback from pilot testing. Our pilot studies involved interviews with three experts in MR design and HCI, each with at least 5 years of experience in their respective domains. They evaluated the fluency and usability of early design iterations. Based on their input, we revised the timeline in *Sequential Multimodal*. We also refined the mapping between modalities and information blocks in *Parallel Multimodal*. These designs represent our best attempt to identify plausible, well-reasoned strategies for multimodal coordination, though alternative instantiations following the same conceptual structure may also be viable. Concretely instantiating each strategy is necessary because abstract notions of parallel or sequential multimodal expression do not specify which modality handles which information block—*object*, *point*, *time*, *motion*, or *interaction*—and in what temporal flow. The structured analysis provides this specification, yielding reproducible expression flows that can be held constant across participants and systematically compared in a controlled experiment.

4.1. Parallel Multimodal Design

The *parallel multimodal* approach expresses dynamic events by leveraging the simultaneous capabilities of multiple modalities. This design directly reflects the inherent parallelism of dynamic events—where several processes often unfold concurrently—by enabling users to convey different information blocks through distinct modalities simultaneously (Kendon, 2004; Clark, 2003).

4.1.1. Alignment of Modalities and Information Blocks

We mapped each information block to the channel whose processing characteristics best match the type of information it carries (Wickens, 2002), distributing cognitive demand across separable resource pools. Speech handles conceptual and categorical information—particularly object initialization and interaction description—where verbal code compatibility is highest (Pinker, 2003; Clark, 1996). For example, in a car collision scenario, speech would be used to name the objects involved ("Car A and Car B") and describe their interaction ("they collide"). Spatial information, including points and trajectory paths, is assigned to gesture and gaze, which leverage spatial processing and manual response modalities to encode multiple spatial relations in parallel, bypassing the serial output constraint of the verbal-vocal channel (Kendon, 2004; Clark, 2003).

Within this allocation, gaze handles *point* indication via fixation, providing a stable and precise spatial reference that naturally aligns with attentional focus (Posner, 1980). While saccadic eye movements can convey direction, they are difficult to control precisely and thus less suited for trajectory expression. Instead, gesture conveys *motion* trajectories and action demonstrations through the spatial-manual channel, enabling users to externalize movement patterns in three-dimensional space (Alibali, 2005). This spatial-manual encoding also captures *time* structure embedded in individual *motion* segments, while speech serves as a fallback for expressing the sequencing of multiple *motions*. Together, these allocations distribute communicative load across channels operating within their regions of highest code compatibility, reducing the concentration of demand on the verbal-vocal channel alone. To make this concrete, consider describing a car moving along a curved path through an intersection. A user would simultaneously speak the object name ("Car A"), fixate gaze on the car's starting position, and trace the curved trajectory through the intersection with a hand gesture. These three modalities operate in parallel, each handling the information type it is best suited

to express.

4.1.2. Parallel Flow for Dynamic Event Expression

As illustrated in Fig. 2, the parallel expression flow enables concurrent communication across multiple channels, with modalities operating simultaneously.

For single-object events (Fig. 2a), users concurrently provide speech for *object* initialization, gaze fixation for *point* indication, and gesture for *motion* demonstration, producing a temporally compressed, naturally integrated expression. For two-object events (Fig. 2b), the flow supports both synchronous and asynchronous scenarios by enabling multimodal expression for each *object*. The process begins with speech for *object* initialization, while gaze fixation specifies the starting *points* for both objects—performed in sequence when the objects are not collocated. Users then demonstrate motion trajectories via gesture—simultaneously for synchronous events or with overlapping timing for asynchronous ones. When a gesture alone cannot fully convey *motion* complexity, speech supplements with additional description. The flow concludes by treating the endpoint of each demonstrated trajectory as the object's ending *point*: in synchronous events, both endpoints are produced simultaneously through the two-handed deictic gesture, whereas in asynchronous events each endpoint is captured when the corresponding object's motion gesture ends. For two-object events with interaction (Fig. 2c), the flow extends the above by inserting an *interaction* phase between the initial motion and final *point* indication. Here, gesture demonstrates contact or collision while speech describes the *interaction* type. Its implementation is shown in Fig. 4c for the user study.

4.2. Structured Sequential Multimodal Design

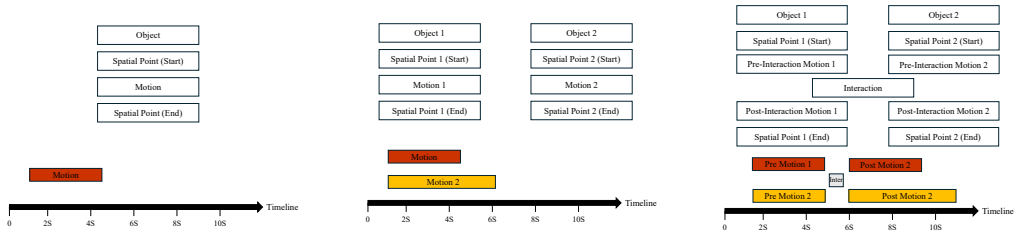
While parallel input is well-suited to dynamic events, our pilot studies revealed that users do not always find it easy to execute, particularly as event complexity increases. When describing multi-object events with interactions, users tended to think through the event step by step rather than expressing all information simultaneously. This observation motivated a second design: we scaffolded the sequential, speech-primary expression process to make it more structured and manageable.

4.2.1. Information Blocks Visualization

Building on our structured analysis, we designed an MR interface to support the sequential expression of dynamic events, with the primary goal of

reducing the cognitive effort required to organize and articulate event details. A key challenge with unstructured speech-only expression is that users frequently omit spatial or temporal details, particularly when the event involves multiple objects and interaction phases. To address this, the interface visually presents all necessary information blocks as scaffolding guidance, externalizing the expression sequence (Sweller, 1994; Sweller et al., 1998) and serving as persistent reminders of what remains to be described. This allows speech to remain the primary modality, while visual cues prompt users through each block in turn, helping to prevent omissions without disrupting the natural flow of speech.

As shown in Fig. 3, the interface adapts to event complexity. For single-object events (Fig. 3a), it presents the basic blocks: *Object*, *Point* (Start), *Motion*, and *Point* (End). For two-object events (Fig. 3b), the blocks are split by object to maintain organization. For two-object events with interaction (Fig. 3c), the motions are divided into pre-*interaction* and post-*interaction* segments for each object, separated by the *Interaction* block. This structure emerged from pilot studies as a configuration users found comfortable and cognitively manageable. All blocks with white backgrounds serve only as visual prompts, functioning as passive reminders rather than interactive elements. The sole exception is the *motion* bar, shown with red or yellow highlights on the timeline, which incorporates gesture-based controls as described next.



(a) UI for single-object event.

(b) UI for two-object event.

(c) UI for two-object event with interaction.

Figure 3: Interfaces for *Sequential Multimodal* expression of (a) single-object, (b) two-object, and (c) two-object events with causal interaction. Information-block panels guide the expression sequence. On the timeline, horizontal position encodes motion order, overlapping bars indicate simultaneity, and bar length represents duration. In (c), the interaction block separates and anchors the pre- and post-interaction motion phases.

4.2.2. Temporal Coordination Through Gestural Manipulation

Speech has inherent limitations in describing precise temporal relationships, such as simultaneity or relative duration (Jackendoff, 2003; Levelt, 1993). To address these limitations, the interface integrates a timeline with *motion* bars beneath the information blocks. Users rely mainly on speech for event description while reserving gestural manipulation specifically for temporal coordination—a division of labor that maintains the expressive strengths of speech while compensating for its temporal constraints.

Each *motion* bar represents a *motion* segment along the temporal dimension, which users can adjust via hand gestures to specify timing, sequence, and duration—dimensions that are difficult to articulate verbally (McNeill, 1992). As illustrated in Fig. 3, dragging the bars along the timeline establishes relative temporal relations: overlapping bars indicate simultaneity, horizontal position (x-axis) encodes motion order, and bar length represents duration. In events involving *interaction*, the timeline is divided into pre- and post-*interaction* phases. The *interaction* block (grey background) acts as a fixed temporal anchor whose position along the timeline can be shifted to mark the point of *interaction*. The implementation of this design in MR is shown in Fig. 4b for the user study.

Taken together, the two designs represent complementary strategies for expressing dynamic events in MR. The *parallel multimodal* emphasizes the simultaneous coordination of modalities, capitalizing on the natural concurrency of dynamic events. The structured *sequential multimodal*, by contrast, prioritizes guided, speech-first articulation supported by visual scaffolding and gestural timeline manipulation. Both approaches were found to be promising in pilot studies, and implementing both allows us to systematically examine how different multimodal coordination strategies affect the speed and the perceived ease and completeness of dynamic event expression.

5. User Study: Comparing Different Expression *Modes*

The previous sections presented a structured analysis that decomposes dynamic events (Section 3) to address RQ1 and developed proof-of-concept prototypes for two multimodal expression strategies (Section 4). This analysis predicts improved efficiency from distributing information across separable resource pools, but it also suggests that synchronization costs in parallel production may rise as coordination demands increase, potentially offsetting the benefits of parallelism. To test this prediction empirically and address

Table 1: Demographic Summary of Participants.

Gender	**Age**	**Education Level**	**MR Experience**	**Self-reported Expression Skill (1-7)**
Male (14)	$M = 26.25$	Bachelor (2)	Never (4)	$M = 3.96$
Female (10)	$SD = 2.72$	Master (11)	< 1 year (11)	$SD = 0.86$
	Range: 21–31	PhD (11)	1–3 years (5), > 3 years (4)	

RQ2, we conducted a controlled within-subjects study comparing the three expression *modes* across the four complexity levels from our structured analysis.

5.1. Participants

We recruited 24 participants (10 female, 14 male; mean age = 26.25 years, SD = 2.72) from the university community. All had normal or corrected-to-normal vision, no color deficiencies, normal speech, and normal hand dexterity. Four participants reported no prior MR experience (Table 1). All were native speakers of the experiment language. The study received institutional review board approval, and participants received compensation of approximately USD 6.50 per hour, aligning with local standards.

5.2. Apparatus

The study took place in a controlled laboratory to minimize distractions. We used the Meta Quest Pro[2] as the MR platform due to its gaze and gesture support. It features a binocular display with a resolution of 1800×1920 pixels per eye at a refresh rate of 90 Hz. The field of view (FOV) is 106° horizontally and 95.57° vertically, with a binocular overlap of 79.72°. The inter-pupillary distance (IPD) range is 55–75 mm.

The experiment ran in Unity3D (2022.3.43f1), displaying dynamic events (e.g., ball movements) in MR while capturing user expressions. Each ball had a diameter of 10 cm. The Meta XR All-in-One SDK[3] and Oculus XR Plugin[4] enabled hand and gaze interactions. A desktop PC (AMD Ryzen 5

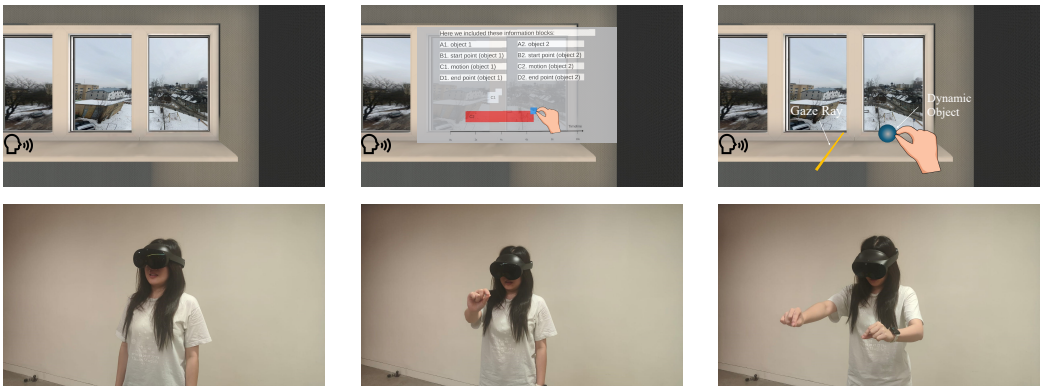
[2] `https://www.meta.com/quest/quest-pro/`

[3] `https://developers.meta.com/horizon/downloads/package/meta-xr-sdk-all-in-one-upm/`

[4] `https://docs.unity3d.com/2022.3/Documentation/Manual/com.unity.xr.oculus.html`

3600 CPU, NVIDIA GeForce RTX 4060 Ti GPU) powered the system via Meta Quest Link, allowing real-time experimenter monitoring.

We did not measure tracking accuracy in our setup, but published evaluations provide baseline estimates. For the Meta Quest Pro eye tracker, Wei et al. (2023) reported an average accuracy of 1.65° and a precision of 0.70° within a 15° field during free head movement. For the Quest Pro's hand tracking, Godden et al. (2025) reported an average positional error of 1.22 cm in the best-performing conditions and a latency that varied widely, from 15.8 ms to 229.2 ms.



(a) *Sequential Speech.* (b) *Sequential Multimodal.* (c) *Parallel Multimodal.*

Figure 4: The three study conditions. The top row shows their implementation in MR, and the bottom row shows a participant performing the corresponding expression tasks. (a) *Sequential Speech* (i.e., unimodal speech) uses free-form spoken description; (b) *Sequential Multimodal* uses structured information-block prompts and gestural timeline manipulation; and (c) *Parallel Multimodal* combines speech, gesture, and gaze concurrently.

5.3. Events

To assess how expression *modes* handle varying cognitive and synchronization demands, we instantiated our structured analysis of dynamic event complexity (Section 3.3) as four *events* (Fig. 5). These progressively varied *objects*, *points*, *times*, *motions*, and *interactions* to represent the four complexity levels.

- *Event 1*: Single-Ball Event — A ball falls and moves to a location.
- *Event 2*: Synchronous Two-Ball Event — Two balls move simultaneously with identical start and end times but to different locations.

- *Event 3*: Asynchronous Two-Ball Event — Two balls start simultaneously but end at different times, moving to different locations.
- *Event 4*: Collision Event — Two balls move synchronously until colliding, then asynchronously.

These four events do not vary along a single continuum. Instead, each event adds a specific demand relative to the previous one: *Event 2* adds a second object, *Event 3* introduces asynchronous timing, and *Event 4* adds a causal interaction. We use *event complexity* as a shorthand for this ordered increase in coordination demand. These levels are ordered, but they are not on an interval scale. As a result, performance differences between events cannot be isolated to a single factor.

5.4. Study Design

We compared three expression *modes* (*Sequential Speech*, *Sequential Multimodal*, *Parallel Multimodal*) in MR (Fig. 4) using a within-subjects design across the four *events* (Fig. 5). We counterbalanced the expression *modes* using a Latin Square, while *events* always followed an order of increasing complexity. We chose a fixed order because the *events* form a progression of difficulty, whereas the *modes* represent independent conditions. This structure allowed participants to build up skills from simpler to more demanding tasks. To reduce learning effects, participants completed a practice session with a dummy *event* before each test *event*. Furthermore, each *event* type featured three comparable variations, ensuring each *mode* was tested with a unique instance. Even so, because each *event* level occurred at the same point in the session, *event* effects cannot be fully separated from order effects.

5.4.1. Conditions and WoZ Scope

The three conditions were designed to compare expression strategies under a common MR task paradigm, not to benchmark end-to-end recognition systems. In all conditions, participants themselves viewed the event, formulated its description, and produced the required expression; task completion time was demarcated by their spoken "start" and "end" markers.

- *Sequential Speech* (Fig. 4a) was a speech-only baseline: participants verbally described the event, and the system recorded their utterances without semantically interpreting them online.

- *Sequential Multimodal* (Fig. 4b) combined speech with passive information-block prompts and direct gestural manipulation of motion bars on a timeline. Participants used speech to articulate the event while the panels scaffolded the required information and the timeline supported temporal coordination; the spoken content was not semantically recognized or fused with another modality during the trial.

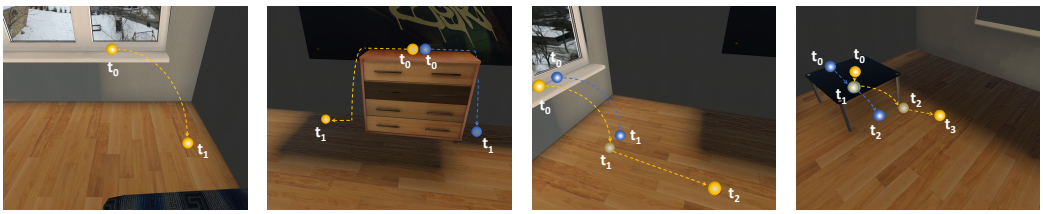
- *Parallel Multimodal* (Fig. 4c) enabled participants to express the event concurrently through speech, gesture, and gaze. Participants used speech to introduce objects and describe interactions, gaze to indicate initial spatial locations, and gestures to demonstrate trajectories and temporal coordination. This design required the system to combine a spoken object reference with a gaze-indicated initial position in real time, for which a reliable pipeline was unavailable. We therefore used a Wizard-of-Oz (WoZ) setup (Dahlbäck et al., 1993) only for this object-instantiation step: when a participant introduced an object verbally (e.g., "A red ball...") while fixating its intended initial position, the experimenter triggered the corresponding object at that location. The wizard did not determine the participant's trajectories, temporal coordination, or interaction description.

This selective use of WoZ held the speech–gaze fusion step in an idealized, error-free design condition, preventing recognition errors and variable fusion latency from being conflated with the user-side cost of coordinating parallel modalities. It was not used to give the parallel condition an engineering-performance advantage over the other conditions, because those designs had no corresponding online semantic-fusion operation to equalize. Accordingly, the comparison estimates the theoretical user-side performance limits of the expression designs.

The conditions also differ in the support that the interface gives: *Sequential Multimodal* offers information-block prompts and a timeline, and *Parallel Multimodal* offers error-free object instantiation. The study therefore compares three complete interaction strategies. It cannot separate the effect of a single modality, or of parallelism itself, from the rest of each design.

5.4.2. Hypotheses

As discussed in Section 4, multimodal approaches offer more intuitive, parallel expression of dynamic events than linear speech-only (unimodal)



(a) Single-ball event. (b) Sync double-ball event. (c) Async double-ball event. (d) Collision event.

Figure 5: Implementation of expression *events* in MR. Temporal labels indicate key event moments: t_0 denotes the initial time, and subsequent labels (t_1, t_2, t_3) denote later transition or end moments in the event sequence. (a) *Event 1*: one ball moves from its initial position at t_0 to its final position at t_1. (b) *Event 2*: two balls start at t_0 and end at t_1, moving simultaneously over the same time interval. (c) *Event 3*: two balls start together at t_0, but one ends at t_1 and the other ends later at t_2. (d) *Event 4*: two balls start at t_0, collide at t_1, and then move apart to their final positions at t_2 and t_3.

methods, though parallel processing has limits and speech provides reliable support. We thus hypothesized:

- *H1*: Multimodal approaches (*Sequential Multimodal*, *Parallel Multimodal*) will enhance the expression of dynamic events compared to *Sequential Speech*. *H1* is an umbrella hypothesis, which we examine separately for efficiency, user satisfaction, and task workload.
- *H2*: The efficiency of *Parallel Multimodal* over *Sequential Multimodal* will increase with event complexity.
- *H3*: The user satisfaction advantage of *Parallel Multimodal* over *Sequential Multimodal* will diminish as event complexity increases.
- *H4*: Perceived task load will be lower when using multimodal approaches (*Sequential Multimodal*, *Parallel Multimodal*).

5.4.3. Measures

We assessed performance, expression capabilities, and user experience objectively and subjectively (Table 2), drawing from prior work. Objective measures captured ***expression efficiency***: *Task Completion Time* and *Word Count*. Subjective measures assessed ***user satisfaction*** with self-reported (1–7 Likert) ratings of *Perceived Understanding*, *Completeness*, *Confidence*, *Comfort*, *Complexity*, and *Preference*; and ***task workload*** with Raw NASA

Task Load Index (*RTLX*) (Hart, 2006) (six subscales, each rated from 1 to 10, with Performance reverse-coded; the overall score is their mean multiplied by 10).

5.5. Procedure

After briefing and consent, participants put on the MR headset, calibrated gaze and gesture tracking, and familiarized themselves with the apparatus. They then completed expression tasks in counterbalanced *mode* order, progressing through *Event 1–4* sequentially. To offset learning from increasing complexity, a practice session with a dummy event of matching complexity preceded each event; formal trials began only after participant confidence and experimenter verification.

Each session involved 12 expression tasks (4 *events* $\times$ 3 *modes*). Participants viewed each event three times in MR, then expressed it using the assigned *mode*—starting with "start" and ending with "end" to demarcate task duration. During each trial, the event animation played automatically three times in a row. Each playback lasted about three seconds, and participants could not pause or replay the animation. During expression, participants described the event from memory. The MR room remained visible, but the animated balls disappeared. Post-task, they removed the headset, completed questionnaires, and took a minimum 5-minute break.

Finally, participants viewed multi-object (3–4) events, followed by an open-ended interview on *mode* preferences. Sessions were audio- and video-recorded for analysis and lasted approximately two hours.

5.6. Analysis

Audio recordings were transcribed using an online speech-to-text service to calculate word counts. Uniform participant language proficiency (Section 5.1) controlled for ability, omitting it as a covariate. As a preliminary descriptive audit, we constructed an event-specific checklist of the information-block instances prescribed by each formal trial, following the five-block framework in Section 3.1. We then reviewed trial records against this checklist to measure block completion—the proportion of required information blocks expressed per event. This audit assesses coverage of prescribed information blocks, not system interpretation accuracy. In the two multimodal conditions, each block had a dedicated interaction step, making coverage complete by design. In *Sequential Speech*, we judged block coverage by reviewing the transcripts. Because the evaluation criteria differed across

Table 2: Measures and operational definitions. † indicates measures where lower values are better; ‡ indicates measures where values need to be interpreted with comparison with other measures; for all others, higher values are better.

Aspect	**Measure**	**Definition/Operationalization**
Expression Efficiency (Oviatt, 1999; Brennan and Clark, 1996)	*Task Completion Time*†	Time to complete the expression task
	Word Count‡	Number of words used through speech during the expression
User Satisfaction (Clark and Brennan, 1991; Hasson et al., 2018; Oviatt et al., 2004, 2017)	*Perceived Understanding* (1–7)	"I believe the computer clearly understood my expression of the events."
	Completeness (1-7)	"I believe my expression of the event in this task was complete and covered necessary details."
	Confidence (1-7)	"I am confident that I expressed the event clearly in this task."
	Comfort (1-7)	"I felt comfortable and at ease while expressing the event during this task."
	Complexity (1-7)†	"I felt this way of expressing the event was complex for this task."
	Preference (1-7)	"Overall, I feel positive about this way of expressing the event in this task."
Task Workload (Hart, 2006)	*RTLX* (10–100)†	6 items, each rated 1–10 (Performance reverse-coded); overall score = mean × 10 (Hart, 2006)

conditions, we report this audit purely as a descriptive measure. It does not indicate expression correctness, and we do not use it to support efficiency comparisons.

We used two-way repeated-measures ANOVA with *Event 1–4* and *Mode* as within-subjects factors to examine complexity and modality effects. Normality was checked using Shapiro–Wilk tests and sphericity using Mauchly's test. When sphericity was violated, we report Greenhouse–Geisser-adjusted degrees of freedom and p values. Effect sizes are reported as partial η^2 with two-sided 95% confidence intervals derived from the noncentral F distribution.

We followed significant main effects with pairwise comparisons among the three *modes* within each *event*. For these post-hoc tests, the three pairwise contrasts within a single *event* and outcome formed one Bonferroni correction family. For the six *RTLX* components, omnibus p-values were additionally Bonferroni-adjusted across the six subscales separately for *Mode*, *Event* , and *Event* $\times$ *Mode* effect families; we analyzed the overall *RTLX* score separately. Because each of the six subjective measures uses a single 7-point item, we interpret them cautiously as general indicators of user perception.

An interaction effect indicates that *mode* differences change across *events*, but it does not reveal the direction of change. Therefore, we directly tested the directional predictions of *H2* and *H3*. For each participant and *event*, we calculated the difference D between the two multimodal modes (*Parallel Multimodal* minus *Sequential Multimodal*). Using these four differences, we calculated a linear trend score, $C = -3D_1 - D_2 + D_3 + 3D_4$, and evaluated whether its mean differed from zero using a two-sided one-sample t-test ($df = 23$). We report $b = C/10$, representing the average change in difference per event level. The *event* sequence (1–4) specifies only the contrast order; we do not assume equal complexity intervals between events. The estimate b summarizes the trend across the ordered event levels. *Task Completion Time* and *Word Count* formed the *H2* family (Bonferroni factor 2). *Preference*, *Perceived Understanding*, and *Completeness* formed the *H3* family (Bonferroni factor 3). The outcomes in these two families are the primary outcomes; the remaining ratings and the *RTLX* scores are secondary. For each event, we report mean differences with unadjusted 95% CIs and standardized effect sizes (d_z). To ensure robust conclusions regardless of normality, we repeated these trend tests using Wilcoxon signed-rank tests on individual slopes. These directional trend tests are post hoc analyses.

We conducted three supplementary checks. First, linear mixed-effects

models included fixed effects of *Mode*, *Event* , and *Event* × *Mode*, a participant-level random intercept, and uncorrelated participant-level random slopes for *Mode* and *Event* . Denominator degrees of freedom were estimated using the Satterthwaite approximation and verified with the Kenward–Roger approximation. Second, Friedman tests compared the three *modes* within each *event* for the six single-item 7-point measures; corresponding Wilcoxon signed-rank follow-ups used the same within-event Bonferroni family of three design contrasts. Third, for each *event*, we reported the paired mean differences between the two multimodal *modes* with confidence intervals, standardized effect sizes, and a sensitivity analysis based on our sample size. Appendix A.1 contains the full mixed-effects, non-parametric, and per-event precision results.

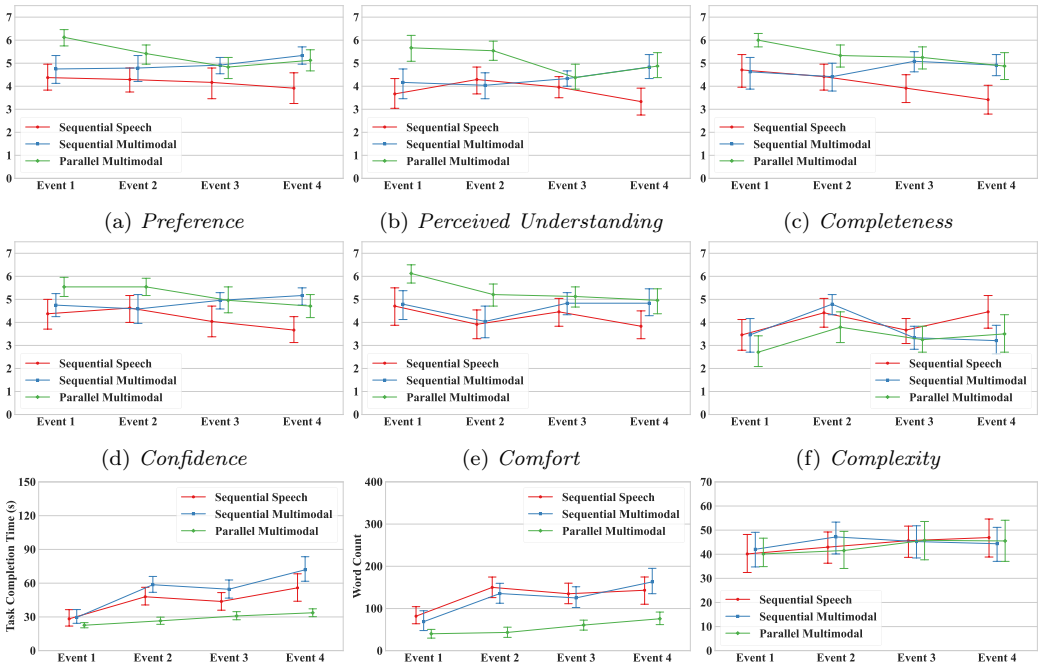
Finally, we transcribed interview recordings directly using online tools and verified them manually against the original audio. We include these participant quotations to illustrate user experiences rather than as part of a systematic qualitative analysis.

5.7. Results

Fig. 6 illustrates the objective and subjective measures across the four *events*. This section reports the main effects and *Event* × *Mode* interaction effects from our statistical analysis, followed by a discussion of how these results relate to our hypotheses. We begin with an overview of user preferences and expression quality before examining the trade-offs in efficiency, satisfaction, and task workload.

5.7.1. Preliminary Information-Block Coverage Audit

As a preliminary check before interpreting the main subjective outcomes, we examined whether participants completed the event-specific required information blocks. Both structured multimodal conditions achieved complete required-block coverage (100%) in all formal trials, by design: their trained interaction flows explicitly elicited each prescribed block. In contrast, *Sequential Speech* achieved an overall coverage of 96.8%, judged from transcripts. Its omissions were concentrated in temporal information: participants sometimes described spatial information but did not subsequently specify the relevant timing, particularly when events required more complex spatiotemporal relations. Because the multimodal interaction flows guaranteed full coverage, this audit cannot distinguish between the two multimodal conditions. It shows only that participants rarely omitted required



(a) *Preference* (b) *Perceived Understanding* (c) *Completeness*

(d) *Confidence* (e) *Comfort* (f) *Complexity*

(g) *Task Completion Time* (h) *Word Count* (i) *RTLX*

Figure 6: User study results across events. Each sub-figure shows average ratings; error bars denote 95% confidence intervals.

information in *Sequential Speech*. Furthermore, this audit does not capture differences in descriptive granularity (e.g., saying "*50 cm behind the sofa*" versus "*next to the sofa*" versus pointing directly). We additionally compared each expression with the corresponding event animation at the level of coarse content. All expressed motions were consistent with the presented *events* (e.g., in their direction of movement). Fine-grained properties, such as exact positions, distances, durations, and relative timing, could not be assessed reliably, for the reason illustrated above. The efficiency results reported below therefore characterize how rapidly and concisely participants conveyed the coarse content of the *events*.

5.7.2. Overall Preferences and Perceived Expression Quality

Across all events (Fig. 6a), multimodal *modes* were generally preferred over the speech-only baseline (*Sequential Speech*). *Parallel Multimodal* emerged as the strongest overall option, particularly for *Event 1* and *Event 2*, where it yielded high *user satisfaction* and efficiency (with error-free speech–gaze fusion). *Sequential Multimodal* took longer, but for the more coordination-

demanding events it received similar ratings, and its structured prompts helped participants feel that they had described the event completely.

This event-dependent preference is clearly reflected in the subjective ratings. For the simplest event (*Event 1*), *Parallel Multimodal* was rated significantly higher than *Sequential Multimodal* in *Preference* ($p_{bonf} = .020$), *Perceived Understanding* ($p_{bonf} = .019$), and *Completeness* ($p_{bonf} = .012$). From the asynchronous event (*Event 3*) onward, however, we no longer found these advantages: at *Event 3* and *Event 4*, the differences in all three measures were close to zero, and their 95% CIs included zero (Table 3). Finding no difference does not mean that the two *modes* are equivalent. At *Event 4*, the CIs still allow differences of up to 0.87 points on the 7-point scale. *Comfort* ratings looked similar (Fig. 6e), but we did not test their trend.

Significant *Event* $\times$ *Mode* interaction effects were found for *Preference* ($F(3.18, 73.04) = 4.13, p = .008, \eta_p^2 = .152$, 95% CI [.013, .287]), *Perceived Understanding* ($F(3.16, 72.65) = 5.29, p = .002, \eta_p^2 = .187$, 95% CI [.033, .326]), and *Completeness* ($F(3.81, 87.73) = 5.28, p < .001, \eta_p^2 = .187$, 95% CI [.040, .311]). To determine the direction of these changes, we tested linear trends directly (Table 3). The *Parallel Multimodal*$-$*Sequential Multimodal* difference decreased linearly across events for *Preference* ($b = -0.55$ points per event level, 95% CI [-0.94, -0.16], $t(23) = -2.89$, $p_{bonf} = .025$, $d_z = -0.59$) and *Completeness* ($b = -0.50$, 95% CI [-0.88, -0.12], $t(23) = -2.75$, $p_{bonf} = .034$, $d_z = -0.56$). *Perceived Understanding* showed a similar declining trend, though it was not significant after correction ($b = -0.58$, 95% CI [-1.08, -0.08], $t(23) = -2.41$, $p = .025$, $p_{bonf} = .074$, $d_z = -0.49$). Wilcoxon signed-rank tests on each participant's slope (Appendix A.1) yielded identical corrected results. The decline was not gradual. Participants rated *Parallel Multimodal* higher at *Event 1*, as well as for *Perceived Understanding* and *Completeness* at *Event 2*. However, differences dropped close to zero from *Event 3* onward. Some participants noted that the cognitive and physical demands of managing parallel inputs eventually offset the *mode's* benefits. As P4 articulated: *"It felt like my hands and mouth were fighting for the same brain space... the moment I had to explain the collision logic, my hands just froze."*

Despite its slower pace, *Sequential Multimodal* produced notable gains in perceived expression quality for events with higher coordination demands. In *Event 4*, it significantly outperformed *Sequential Speech* in *Completeness* ($p_{bonf} = .001$), *Perceived Understanding* ($p_{bonf} = .003$), *Confidence* ($p_{bonf} = .003$), and *Preference* ($p_{bonf} = .002$). Some participants attributed this to

the scaffolding structure of the sequential/stepwise panels. P11 explained: *"The step-by-step prompts acted as an external memory aid, ensuring I didn't miss the small details that were otherwise lost in the flow of free speech."*

For events with higher coordination demands, *Sequential Multimodal* traded speed for perceived quality. Although slower, participants rated it higher than *Sequential Speech*, and its ratings did not differ significantly from *Parallel Multimodal*.

Comparing the two multimodal *modes* suggests that perceived expression quality depends on event coordination demands. For simple events, *Parallel Multimodal* produced higher perceived quality and preference, likely because participants could express spatial and temporal information naturally without interrupting speech. However, for asynchronous and collision *events*, the two multimodal modes showed no significant differences in *Completeness*, *Perceived Understanding*, or *Preference*. The stepwise structure may ease the coordination demands of parallel input, though our data cannot confirm this mechanism.

Finally, *Parallel Multimodal* uniquely reduced perceived task *Complexity* in *Event 4* ($p_{bonf} = .017$) and increased *Comfort* across several events ($p_{bonf} < .05$). Some participants appreciated the intuitive nature of spatial interaction in MR, which allowed them to offload verbal descriptions onto physical gestures. As P3 noted, *"Using my hands to show the path while talking felt much more natural; it removed the need to find the right words."*

5.7.3. Task Completion Time and Efficiency

The repeated-measures ANOVA revealed significant *Event* $\times$ *Mode* interaction effects for both *Task Completion Time* ($F(3.39, 77.97) = 8.69, p < .001, \eta_p^2 = .274$, 95% CI [.102, .410]) and *Word Count* ($F(6, 138) = 6.12, p < .001, \eta_p^2 = .210$, 95% CI [.077, .305]), indicating that the relative efficiency of the *modes* differed across the tested event progression.

Compared to the *Sequential Speech*, *Parallel Multimodal* was consistently faster from *Event 2* through *Event 4* (all $p_{bonf} \leq .012$) and required significantly fewer spoken words at every *event* (all $p_{bonf} \leq .002$). These results indicate that participants used gestures and gaze to take over part of what they would otherwise have said, and so finished their expressions faster and with fewer words. The results describe speed and conciseness, not the precision of the expressed details, and they were obtained with error-free speech–gaze fusion in *Parallel Multimodal*.

Across *events*, the two multimodal *modes* differed markedly in efficiency.

Table 3: Differences between the two multimodal *modes* (*Parallel Multimodal* minus *Sequential Multimodal*) in each *event*, and their linear trend across *events* ($N = 24$). Each cell shows the mean difference and, below it, the unadjusted 95% CI. For time and words, negative values mean that *Parallel Multimodal* was faster or used fewer words. For the 7-point ratings, positive values mean that *Parallel Multimodal* was rated higher. The trend b is the average change in the difference per event level (contrast weights $-3, -1, 1, 3$, divided by 10). p_{bonf} is Bonferroni-adjusted within the *H2* family (factor 2) or the *H3* family (factor 3). Table A.7 gives the condition means, the time and word ratios, and d_z.

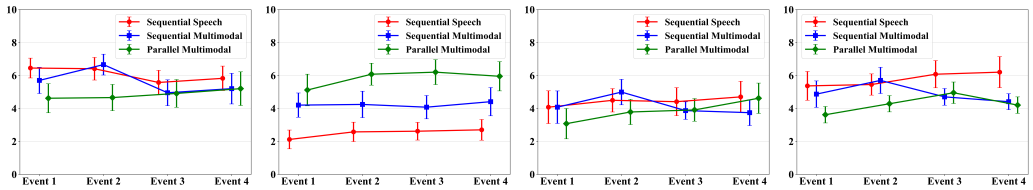
Outcome	**Event 1**	**Event 2**	**Event 3**	**Event 4**	**Trend** b	p_{bonf}
H2 family						
Time (s)	−6.95 [−13.65, −0.25]	−32.07 [−40.16, −23.98]	−23.75 [−33.17, −14.33]	−38.19 [−48.76, −27.63]	−8.54 [−11.52, −5.56]	< .001
Words	−28.6 [−55.8, −1.5]	−92.1 [−121.9, −62.3]	−64.3 [−93.4, −35.2]	−87.3 [−122.8, −51.9]	−14.83 [−24.16, −5.51]	.006
H3 family						
Preference	1.38 [0.42, 2.33]	0.63 [−0.07, 1.32]	−0.08 [−0.58, 0.41]	−0.21 [−0.84, 0.43]	−0.55 [−0.94, −0.16]	.025
Understanding	1.50 [0.47, 2.53]	1.50 [0.88, 2.12]	0.04 [−0.49, 0.58]	0.04 [−0.79, 0.87]	−0.58 [−1.08, −0.08]	.074
Completeness	1.38 [0.48, 2.27]	0.92 [0.11, 1.72]	0.17 [−0.31, 0.64]	−0.04 [−0.56, 0.48]	−0.50 [−0.88, −0.12]	.034

Main effects of *mode* were significant for both *Task Completion Time* ($F(2, 46) = 32.91, p < .001, \eta_p^2 = .589$, 95% CI [.392, .710]) and *Word Count* ($F(1.46, 33.47) = 34.42, p < .001, \eta_p^2 = .599$, 95% CI [.368, .734]), among the largest effects in the study. Post-hoc comparisons showed that this advantage held for both completion time and word count at *Event 2–Event 4* (all $p_{bonf} < .001$), but not at *Event 1* (time: $p_{bonf} = .128$; words: $p_{bonf} = .119$). We report the per-event differences and their trend in Table 3 rather than repeating them here.

H2 predicts that this advantage grows with event level. To test this, we examined the *Parallel Multimodal*−*Sequential Multimodal* differences (Table 3). *Parallel Multimodal* was faster on average at every level: by 6.95 s at *Event 1*, 32.07 s at *Event 2*, 23.75 s at *Event 3*, and 38.19 s at *Event 4* ($d_z = -0.44, -1.67, -1.06$, and -1.53), and the difference grew linearly across event levels ($b = -8.54$ s per level, 95% CI [−11.52, −5.56], $t(23) = -5.92$, $p_{bonf} < .001$, $d_z = -1.21$). Word count followed the same pattern ($b = -14.83$ words per level, 95% CI [−24.16, −5.51], $t(23) = -3.29$, $p_{bonf} = .006$, $d_z = -0.67$). Wilcoxon signed-rank tests on each participant's

slope yielded identical conclusions. However, this efficiency gain was not strictly monotonic at every step. Time savings were smaller at *Event 3* than at *Event 2*, because *Sequential Multimodal* became faster (from $M = 58.62$ s to 54.63 s) while *Parallel Multimodal* became slower (from $M = 26.55$ s to 30.88 s). In relative terms, *Parallel Multimodal* needed 76.5% of the *Sequential Multimodal* time at *Event 1*, compared to 45.3% at *Event 2*, 56.5% at *Event 3*, and 46.9% at *Event 4*. Thus, relative time savings grew from *Event 1* to *Event 2*, but stabilized after that.

Conversely, *Sequential Multimodal* was significantly slower than *Sequential Speech* in *Event 2* ($p_{bonf} = .017$) and *Event 4* ($p_{bonf} = .031$), but not in *Event 1* or *Event 3* ($p_{bonf} = .050$), indicating that adding structured multimodal panels can introduce a measurable time cost. Overall, the efficiency results partially support *H2*, *The efficiency of Parallel Multimodal over Sequential Multimodal will increase with event complexity.* The absolute time difference between *Parallel Multimodal* and *Sequential Multimodal* showed a significant linear trend across the ordered event levels; this increase was non-monotonic, and the relative advantage plateaued after *Event 2*. Because speech–gaze fusion was error-free, these results show how efficiently users can coordinate the modalities when fusion never fails. They do not show how efficient a deployed recognition system would be: real recognition errors, latency, and recovery demands could reduce or otherwise alter the advantage, and we did not test end-to-end recognition. They also clarify the role of *Sequential Multimodal*: in the more coordination-demanding events, it gives up speed in exchange for scaffolding and higher perceived expression quality.



(a) Mental demand. (b) Physical demand. (c) Frustration. (d) Performance.

Figure 7: Results of the Raw NASA Task Load Index (*RTLX*) sub-scales across events. Error bars denote 95% confidence intervals. The *Performance* item was reverse-coded so that higher values indicate poorer perceived performance, consistent with *RTLX* scoring.

5.7.4. Task Workload

An analysis of the overall *RTLX* scores revealed a significant *Event* $\times$ *Mode* interaction effect ($F(2.65, 61.02) = 3.06, p = .041, \eta_p^2 = .117$, 95% CI [.000,

.260]); however, workload did not consistently decrease under the *Parallel Multimodal mode*. We therefore decomposed *RTLX* into its six subdimensions. *Mental Demand*, *Physical Demand*, and *Performance* showed familywise-significant *mode*-related effects, whereas *Frustration* showed an uncorrected interaction that did not survive the across-component correction (Fig. 7). *Temporal Demand* and *Effort* showed no significant main effects or *Event* $\times$ *Mode* interactions (Appendix A.2). We report these design-related sub-dimensions below.

Mental Demand. Mental demand showed a significant main effect of *mode* ($F(1.51, 34.82) = 6.50, p = .007, p_{\text{bonf}} = .045, \eta_p^2 = .220$, 95% CI [.020, .432]). The *Event* $\times$ *Mode* interaction was significant before correction ($F(2.69, 61.89) = 3.61, p = .022, p_{\text{bonf}} = .129, \eta_p^2 = .136$, 95% CI [.002, .282]) but did not survive the across-component Bonferroni correction. We therefore treat the event-specific pattern as exploratory. Within-event comparisons showed lower mental demand for *Parallel Multimodal* than *Sequential Speech* in *Event 1* ($p_{bonf} = .007$) and *Event 2* ($p_{bonf} = .011$), and than *Sequential Multimodal* in *Event 2* ($p_{bonf} < .001$). These differences were absent in *Event 3–Event 4* ($p_{bonf} > .05$), a pattern suggestive of coordination demands becoming comparable across modes in these later events.

Physical Demand. A significantly large main effect of *mode* was observed ($F(1.50, 34.44) = 60.18, p < .001, p_{\text{bonf}} < .001, \eta_p^2 = .724$, 95% CI [.548, .817]). The *Event* $\times$ *Mode* interaction did not reach conventional significance ($F(2.59, 59.49) = 2.13, p = .115, p_{\text{bonf}} = .689, \eta_p^2 = .085$, 95% CI [.000, .219]), indicating that this mode-level difference was relatively stable across *events*. Across all events, physical demand followed a consistent hierarchy: *Parallel Multimodal* ($M \approx 5.8$) > *Sequential Multimodal* ($M \approx 4.2$) > *Sequential Speech* ($M \approx 2.5$), with all pairwise comparisons significant ($p_{bonf} < .001$). The efficiency gains of parallel input, obtained with error-free fusion, may be partly offset by the physical effort required for sustained mid-air gestures and gaze. However, given that participants consistently preferred *Parallel Multimodal* in the less coordination-demanding events, this physical cost was generally deemed acceptable in practice.

Frustration. The *Event* $\times$ *Mode* interaction was significant before correction ($F(3.09, 71.07) = 3.94, p = .011, p_{\text{bonf}} = .066, \eta_p^2 = .146$, 95% CI [.010, .282]) but did not survive the across-component Bonferroni correction. We therefore treat the event-specific pattern as exploratory. In *Event 2*, *Sequential*

Multimodal produced higher frustration ($M = 5.0$) than both *Parallel Multimodal* ($M = 3.8, p_{bonf} = .034$) and *Sequential Speech* ($M = 4.5, p_{bonf} = .022$). This pattern suggests that serialized interface steps may introduce interaction friction for tasks that invite bimanual expression, but it should not be interpreted as confirmatory evidence or as a general rejection of sequential scaffolding across all events.

Performance. Performance also showed both a significant main effect of *mode* ($F(1.40, 32.25) = 9.61, p = .002, p_{\text{bonf}} = .010, \eta_p^2 = .295$, 95% CI [.057, .507]) and a significant *Event* $\times$ *Mode* interaction effect ($F(3.09, 70.96) = 5.02, p = .003, p_{\text{bonf}} = .018, \eta_p^2 = .179$, 95% CI [.027, .320]). Because the original questionnaire item asked participants how successful they felt, we reverse-coded this item for consistency with *RTLX* scoring; lower *Performance* scores therefore indicate better perceived task performance. Post-hoc comparisons showed that *Sequential Speech* led to better perceived performance than *Parallel Multimodal* in *Event 1* ($p_{bonf} = .002$) and *Event 4* ($p_{bonf} = .002$). *Sequential Multimodal* led to better perceived performance than *Parallel Multimodal* in *Event 2* ($p_{bonf} = .023$), while *Sequential Speech* led to better perceived performance than *Sequential Multimodal* in *Event 3* ($p_{bonf} = .025$) and *Event 4* ($p_{bonf} = .014$). These results suggest that perceived task performance does not simply follow efficiency: although *Parallel Multimodal* was faster, participants sometimes felt more successful with speech-based or structured sequential expression, especially when precision and control became important.

Summary. Overall, *Parallel Multimodal* reduced mental demand in *Event 1* and *Event 2* and avoided the frustration spike observed for *Sequential Multimodal* in *Event 2*, but these benefits came at the cost of increased physical demand, and its speed advantage did not translate into perceived success as coordination demands increased. This dynamic represents a workload *redistribution* rather than a strict reduction, aligning with Multiple Resource Theory (Wickens, 2002): offloading information processing across different sensory channels may minimize cross-channel interference without necessarily lowering total resource expenditure. P12 captured this trade-off: *"Parallel input made it easier to get the timing right, but drawing the trajectories precisely in the air was physically more demanding than I expected."*

5.7.5. Summary of Hypotheses.

Table 4 summarizes the findings corresponding to each hypothesis and its support status.

Table 4: Summary of hypotheses, corresponding findings, and support status. Support status for *H2* and *H3* is based on the linear trend contrasts in Table 3.

Hypothesis	**Corresponding findings**	**Support status**
H1: Multimodal approaches (*Sequential Multimodal*, *Parallel Multimodal*) will enhance the expression of dynamic events compared to *Sequential Speech*.	*Parallel Multimodal* was generally faster and more concise than *Sequential Speech* (with error-free speech–gaze fusion), while *Sequential Multimodal* improved perceived expression quality over *Sequential Speech* for *Event 4*. The benefits depended on the *event* and outcome.	Supported, with event-dependent benefits.
H2: The efficiency of *Parallel Multimodal* over *Sequential Multimodal* will increase with event complexity.	With error-free speech–gaze fusion, *Parallel Multimodal* was on average faster and used fewer words than *Sequential Multimodal* in every *event*. The time saving grew linearly ($b = -8.54$ s per level, $p_{bonf} < .001$), though not at every step, and the relative saving was similar from *Event 2* to *Event 4*.	Partially supported.
H3: The user satisfaction advantage of *Parallel Multimodal* over *Sequential Multimodal* will diminish as event complexity increases.	*Parallel Multimodal* was rated higher at *Event 1*–*Event 2*, but the differences were near zero from *Event 3* onward. The downward trend was significant for *Preference* and *Completeness*, but not for *Perceived Understanding* after correction. The *Event 4* results do not show equivalence.	Partially supported.
H4: Perceived task load will be lower when using multimodal approaches (*Sequential Multimodal*, *Parallel Multimodal*).	Multimodal input reduced mental demand in some events, but increased physical demand and did not consistently reduce overall workload.	Partially supported.

6. Discussion

6.1. Performance Trade-offs Across Three Expression Modes

Our user study compared three *modes* for expressing dynamic events in MR: *Sequential Speech*, *Parallel Multimodal*, and *Sequential Multimodal*. The results show that no single approach dominates across all situations—the right choice depends on the coordination demands of the event, and the relationship between efficiency and user experience is more nuanced than a simple "more modalities = better" story. Because the three conditions are complete interaction strategies and all participants completed the events in the same fixed order, the observed differences may also reflect interface support, practice, or fatigue alongside event demands.

For simple events, speech alone was sufficient: users could describe a single moving object without gestural support. For events with two moving objects, *Parallel Multimodal* (combining hand gestures and eye gaze with speech) completed the task faster under error-free speech–gaze fusion, but also introduced greater physical strain and did not always yield proportional gains in satisfaction. For the collision event, with interacting objects and causal logic, *Sequential Multimodal*, which guides users step by step, received higher ratings of completeness and satisfaction than *Sequential Speech*, and ratings that did not differ significantly from those of *Parallel Multimodal*, despite being the slowest approach.

Understanding why these trade-offs occur requires examining what each approach demands of users, both cognitively and physically. The next two sections unpack this, first qualitatively, then through a formal model.

6.2. Efficiency, Workload, and the Cost of Coordination

Our data surface a persistent paradox: *Parallel Multimodal* was the fastest approach under error-free speech–gaze fusion, yet users did not consistently rate it as the most satisfying or least demanding. This tension may point to two cognitive mechanisms operating beneath task performance: the cost of maintaining transient signals in working memory, and a physical limit on how many objects two hands can simultaneously track. Section 6.3 uses these mechanisms to motivate an illustrative analysis inspired by the Keystroke-Level Model (KLM) (Card et al., 1980) of how coordination costs could affect parallel expression as coordination demands increase.

6.2.1. Mechanism-Related Concepts

We use the following concepts to interpret the observed trade-offs. They relate the resource-distribution benefit of parallelism described by Multiple Resource Theory to the coordination demands required to sustain parallel output (Wickens, 2002, 2008; Salvucci and Taatgen, 2008). First, *synchronization cost* denotes the user-side cognitive and motor effort required to plan, initiate, temporally align, and monitor complementary modality streams so that they jointly convey one coherent dynamic event. It is a theoretical mechanism, not an independently measured behavioral variable.

To keep what we observed apart from what we infer, we distinguish three levels. *Observed:* we measured completion time, word count, the six subjective ratings, and *RTLX* workload, which showed a large increase in physical demand and event-dependent differences in mental demand. *Inferred:* synchronization cost itself was not measured. We use it as one possible explanation for why the subjective advantages of *Parallel Multimodal* declined while its time advantage remained. Because speech–gaze fusion in *Parallel Multimodal* was error-free, recognition errors cannot explain this decline. Other contributing factors may include physical fatigue from prolonged mid-air gestures, limited gesture experience, UI friction, tracking uncertainty, or the fixed event order. *Measurable in future work:* specific behavioral indicators could make synchronization cost directly testable. Future studies could measure speech–gesture onset lag, speech pauses and disfluencies, interrupted gestures, retraced trajectories, modality switching, error corrections, and gaze shifts away from target objects.

We use *cognitive boundary of parallel expression* to denote the condition-specific regime in which parallel expression no longer yields a reliable subjective advantage over structured sequential expression, despite maintaining a speed advantage. In the present study, this regime is indicated by two results: 1) the subjective advantage of *Parallel Multimodal* over *Sequential Multimodal* declined significantly across events for preference and completeness (though not for perceived understanding after correction), and 2) starting from the asynchronous event, we found no significant differences in these subjective ratings between the two multimodal modes. Because we did not perform equivalence testing, this boundary only marks where we could no longer detect a reliable subjective advantage. It does not establish a universal human threshold, a fixed object limit, or a loss of objective efficiency. The term labels this pattern in the subjective ratings.

We use *multimodal ceiling* to describe the corresponding saturation of marginal user-experience benefit: further parallelization no longer yields a net subjective benefit, possibly because added coordination costs offset the added expressive capacity. It is not an exact crossover point estimated by this study or a fixed limit of human multimodal capacity. All three concepts are therefore interpreted through workload, subjective outcomes, completion time, and qualitative reports as indirect evidence, rather than directly measured variables.

6.2.2. The Transient Information Problem

Because speech and gesture are ephemeral, coherent parallel expression requires users to maintain both streams in working memory to keep them semantically aligned (Baddeley, 2020; Cowan, 2010; Wickens, 2002). This burden is likely to increase with coordination demands as transient information approaches working-memory capacity. P4 (Section 5.7.2) described their hands and mouth as "fighting for the same brain space" until their hands froze when the collision logic had to be explained. We refer to this reported experience as "cognitive jamming". Demonstrating motion relies on visuospatial-motor processing, whereas explaining causal logic is verbal-analytical; these processes can run together at low object counts but may increasingly compete as N and coordination demands rise. Our data fit this account, although they cannot confirm it. The subjective advantage of *Parallel Multimodal* was no longer detected from *Event 3* onward, where the two objects' timing had to be expressed separately. This happened before any causal interaction was added in *Event 4*.

This account may also explain the strengths of *Sequential Multimodal*. Instead of requiring users to hold the entire scene in working memory while simultaneously gesturing and narrating, the *mode's* step-by-step prompts served as external retrieval cues, directing attention to one piece of information at a time (Baddeley, 2020; Kirsh, 2010). This scaffolding may have mitigated omissions that high working-memory load could otherwise cause, resulting in descriptions that participants perceived as more complete (Fig. 6c). However, this completeness came at the cost of fluency: the enforced wait states created a temporal dissonance between users' rapid mental simulations and the *mode's* rigid pacing. P9 described this as knowing where the hands needed to go but being forced to wait for the interface.

6.2.3. The Bimanual Capacity Limit

The second constraint is strictly physical: human bimanual capacity is limited to two hands. For $N \leq 2$, parallel expression maps intuitively onto the body (one hand per object), yielding peak efficiency. Beyond two objects, this 1:1 mapping breaks down; users cannot simultaneously point to three distinct trajectories. To adapt, participants chunked objects into bimanual pairs and processed each pair in separate rounds, as articulated by P7.

This chunking strategy may preserve some time savings, but it may also introduce a regrouping cost at each round boundary. Users may need to pause to recall which objects had been addressed, reassign hand roles, and update their spatial mental model. This overhead may grow with each completed round, because the mental model of processed objects expands and must be actively suppressed to prevent interference (Cowan, 2010). While chunking may preserve some of the time savings (Fig. 6g), the mental effort for resequencing may prevent a corresponding reduction in perceived mental demand (Fig. 7a).

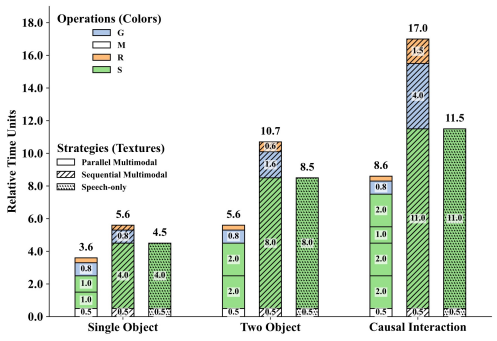
6.2.4. How the Two Mechanisms Combine

Together, these mechanisms motivate a conditional scaling hypothesis: if coordination and regrouping costs increase across repeated bimanual rounds, their accumulation can eventually offset the efficiency gained from parallel execution, whereas the serial baselines grow approximately linearly with the number of objects.

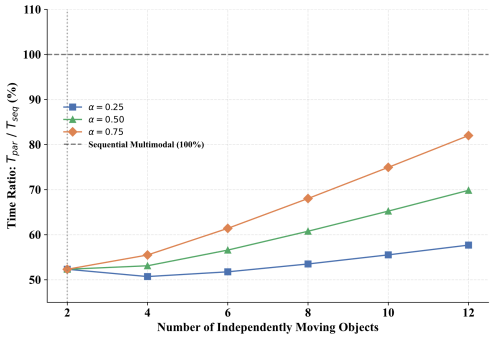
Our formal quantitative trials do not test this scaling hypothesis beyond $N = 2$. *Parallel Multimodal* retains a clear time advantage for two-object events under error-free speech–gaze fusion, while its subjective advantage diminishes as coordination demands increase (Fig. 6a). This pattern is qualitatively consistent with the proposed mechanism, but does not validate a crossover at higher values of N.

The subsequent analysis formalizes transient-information and bimanual-regrouping costs as explicit assumptions. It is intended to make the resulting conditional implications testable in future studies, rather than to substitute for empirical evaluation at higher object counts.

Figure 9 connects these mechanisms to the subsequent KLM-inspired analysis. The parallelism-benefit curve represents the increasing, but bimanually bounded, time savings made possible by overlapping speech, gesture, and gaze, whereas the synchronization-and-regrouping-cost curve represents the coordination overhead that can accumulate as event complexity and the



(a) Micro-level operator composition for simple events ($N \leq 2$).



(b) Illustrative relative completion-time scenarios between *parallel multimodal* and *sequential multimodal* as object number increases.

Figure 8: Keystroke-Level Model (KLM)-inspired operator analysis and illustrative sensitivity scenarios for multimodal dynamic event expression. (a) Micro-level operator composition illustrates how *parallel multimodal* can compress completion time for simple events by overlapping speech, gesture, gaze/head pointing, mental preparation, and system response operators. S: Speaking; G: Gesturing; Z: Gaze/head pointing; M: Mental preparation; R: System response. (b) Macro-level sensitivity scenarios plot the relative completion time $T_{\text{par}}/T_{\text{seq}}$ as the number of independently moving objects N increases. Values below 100% indicate that *parallel multimodal* is faster than *sequential multimodal* under the assumed cost function. Different curves vary the uncalibrated regrouping-cost parameter α, where larger values represent greater assumed overhead from reassigning hands and updating the spatial mental model after each bimanual round. The curves do not estimate an empirical crossover point.

number of bimanual rounds increase. The vertical axis expresses the conceptual relative magnitude of these competing effects rather than a measured score or calibrated KLM unit. Their conditional intersection, C^*, corresponds conceptually to a scenario-specific solution of $T_{\text{par}} = T_{\text{seq}}$. We do not estimate C^* in this study: its position depends on the unmeasured regrouping function $f(k)$ and its parameters, and only illustrates how increasing coordination demands could make sequential scaffolding competitive.

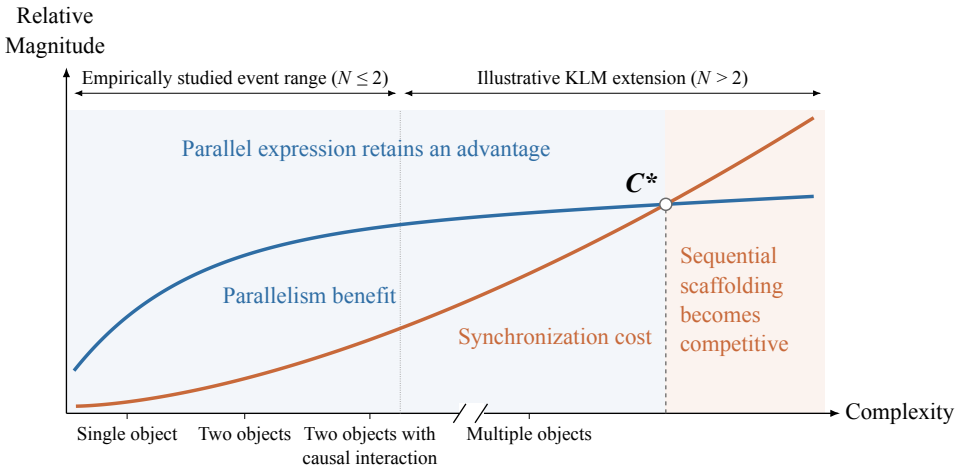


Figure 9: Keystroke-Level Model (KLM)-inspired conceptual model relating event complexity, the benefit of parallel execution, and synchronization and regrouping costs. Parallelism initially provides increasing benefit by overlapping speech, gesture, and gaze, but this benefit is bounded by bimanual capacity. Coordination costs are assumed to rise with causal and temporal demands and with repeated bimanual regrouping. Their intersection, C^*, represents a scenario-dependent crossover at which the parallel advantage is offset and sequential scaffolding can become competitive. The $N \leq 2$ region reflects the study scope; the $N > 2$ region and C^* are illustrative rather than empirically estimated. The break in the horizontal axis marks a long, shortened stretch: with the illustrative parameter values in Appendix B.8, C^* would lie far beyond the tested range. Its position in this figure is purely schematic.

6.3. KLM-Inspired Analysis of Coordination Costs

Building on the possible constraints discussed above—working-memory maintenance and bimanual regrouping costs—we use a KLM-inspired operator analysis with CPM-GOMS parallel-channel logic (John and Gray, 1995; Gray et al., 1993). The analysis is not fitted to the observed data and is not intended to predict absolute completion times. Instead, it provides a

structured, illustrative account of how parallel execution and coordination overhead could trade off under stated assumptions. We therefore present it as a secondary extension of the empirical study, meant to generate hypotheses. It is not an empirical model of scalability, and its outputs beyond the tested events are not findings of this study.

All times are expressed in relative units, anchoring one unit to the duration of one speech chunk ($S = 1.0$). Following the procedure detailed in Appendix B.1, this base unit approximates a short spoken phrase, and the remaining operators are scaled using MR-validated measurements (Cabric et al., 2021). We define the core operators: Mental preparation ($M = 0.5$), Gesture/motion ($G = 0.8$), Gaze/head point ($Z = 0.5$), and System response ($R = 0.3$). Event complexity is parameterized by N, the number of independently moving objects. We use c to denote the number of speech chunks required to describe the basic information for one object and set $c = 4$, corresponding to object identity, start position, motion trajectory, and end position. Each object also requires one gesture operator when the strategy involves trajectory demonstration. *Interaction* events introduce additional speech (c_{int}) and gesture (g_{int}) operators. These operators are abstractions of continuous trajectories and are not direct measurements of participants' actions. The mappings and step-by-step derivations are detailed in Appendix B.1–Appendix B.5. Accordingly, Fig. 8b shows relative, scenario-dependent ratios $T_{\text{par}}/T_{\text{seq}}$, rather than calibrated completion-time predictions.

6.3.1. Illustrative Operator Analysis for Simple Events

Within this illustrative operator composition, speech-only and sequential multimodal strategies process all operators serially, so their completion times grow linearly with N:

$$T_{\text{speech}} = M + cNS = 0.5 + 4N \tag{1}$$

$$T_{\text{seq}} = M + cNS + NG + NR = 0.5 + 5.1N \tag{2}$$

The sequential time exceeds speech-only by 1.1 units per object ($G + R = 0.8 + 0.3$). This differential represents the explicit cost of gesture demonstration (G) and system latency (R), which participants subjectively experienced as disruptive pacing.

Parallel Multimodal overlaps gesture, gaze and verbal execution for each object, lowering the per-object cost while incurring a fixed initial overhead:

$$T_{\text{par}} = (M + G + R) + 2.0N = 1.6 + 2.0N \tag{3}$$

In this abstraction, the fixed term (1.6) represents initial mental preparation and an end-point gesture that occur once regardless of N. The per-object cost of 2.0 (slope) represents concurrent speech and gesture within each object's description. Detailed CPM-GOMS operator compositions for all *modes* are provided in Appendix B.3–Appendix B.5.

For $N = 2$, the illustrative operator composition yields 5.6 relative units—roughly 52% of the sequential scenario (10.7) and 66% of the speech-only scenario (8.5). This *bimanual sweet spot* represents the maximum parallelism assumed in the two-hand model. It is qualitatively consistent with the empirical efficiency advantage in Fig. 6g, but the relative values are not calibrated to observed completion times. Fig. 8a contrasts these compositions: the parallel strategy compresses task duration by concurrently executing operators that sequential methods process serially.

6.3.2. Events with Interaction

For an illustrative collision scenario, we set $c_{\text{int}} = 3$ and $g_{\text{int}} = 3$, raising the serial scenarios to $T_{\text{speech}} = 11.5$ and $T_{\text{seq}} = 17.0$. To represent possible cross-domain interference between spatial demonstration and causal narration—the "cognitive jamming" illustrated by P4's account (Section 6.2)—we assume an additional mental regrouping step ($M_r = 1.0$) between the object-description and interaction-description phases, yielding $T_{\text{par}} = 8.6$. This step is an analytical assumption, not a measured operator.

Under these assumptions, the parallel scenario remains advantageous, but its margin narrows relative to the interaction-free scenario. This illustrates how additional coordination requirements could erode the benefit of parallel execution; it does not validate a scaling trend beyond the tested events.

6.3.3. Illustrative Scaling Scenarios Beyond Bimanual Capacity

The formal quantitative trials did not include more than two independently moving objects. To explore a possible scaling scenario beyond bimanual capacity, we assume that users partition objects into $\lceil N/2 \rceil$ bimanual rounds separated by mental regrouping steps, inspired by P7's reported chunking strategy. Under this assumption, the total parallel completion time becomes:

$$T_{\text{par}}(N) \;=\; M + \left\lfloor \tfrac{N}{2} \right\rfloor T_{\text{r2}} + (N \bmod 2)\, T_{\text{r1}} + \sum_{k=1}^{\lceil N/2 \rceil - 1} M_r(k) + G + R \qquad (4)$$

Here, T_{r2} and T_{r1} are the durations of two-object and single-object parallel rounds, respectively. The regrouping cost following the k-th round is $M_r(k) = t_{M_0} + f(k)$, where t_{M_0} is an assumed base hand-reassignment cost and $f(k)$ is an assumed non-negative, monotonically non-decreasing function. We use this function to represent a possible increase in working-memory maintenance as rounds accumulate (Section 6.2); neither its form nor its parameters were measured in this study. Appendix B.6 provides the corresponding derivation.

For example, $N = 4$ entails two assumed rounds and one regrouping episode; $N = 6$ entails three rounds and two episodes. If $f(k)$ increases sufficiently across these episodes, the accumulated cost can produce a scenario-specific crossover, at which T_{par} exceeds a serial baseline. Its location, and whether it occurs within a practical range of N, depend on the unmeasured function $f(k)$ and its parameters. The sensitivity curves in Fig. 8b therefore illustrate conditional outcomes rather than a universal threshold. The derivations and quadratic example used for these scenarios are provided in Appendix B.7.

6.3.4. Scope and Testable Implications

Our formal quantitative study did not test $N > 2$ independent trajectories, so it cannot validate the scaling function, estimate a crossover point, or establish a minimum of the efficiency ratio. P7's reported chunking strategy and the declining subjective appeal of *Parallel Multimodal* in more coordination-demanding two-object events are qualitatively consistent with the mechanism, but they do not test the proposed growth function. The analysis yields two testable hypotheses for future MR research: (1) if regrouping costs accumulate beyond bimanual capacity, $T_{\text{par}}/T_{\text{seq}}$ should increase; and (2) subjective preference for parallel expression should decline in tandem.

Appendix B.8 varies the uncalibrated regrouping-cost parameter to make this dependence explicit. It shows that changing this assumption substantially changes the scenario-specific crossover. The analysis therefore motivates empirical measurement of coordination costs at higher object counts, rather than establishing a hard scalability ceiling.

6.4. Design Implications

Our findings question the assumption that maximizing concurrent multimodal input inherently improves user experience. The empirical data show diminishing subjective benefits as coordination demands increase, while the KLM-inspired analysis identifies conditions under which additional coordination could offset bandwidth gains. Drawing on these observed trade-offs and conditional mechanisms, we propose three design implications for future MR multimodal expressive systems.

6.4.1. Adaptive Scaffolding for Cognitive Load Management

Current MR interfaces typically enforce a binary choice between free-form parallel input and rigid sequential wizards. However, users' capacity to sustain parallel expression may vary with coordination demands. MR systems can adopt real-time, adaptive scaffolding. Similar to how a navigation system gives turn-by-turn instructions in an unfamiliar area but stays quiet on routine commutes, an MR expressive system should remain unobtrusive during simple events. If the system detects cognitive overload—such as prolonged pauses, hesitations, speech disfluencies, or suspended gestures—it should proactively introduce sequential scaffolding to alleviate working-memory demands. Once cognitive resources are freed, the scaffolding should gracefully recede. This paradigm shifts the system from a rigid modality enforcer to an adaptive cognitive partner.

6.4.2. Supporting Entity Chunking Under High Coordination Demand

If coordination costs accumulate beyond bimanual capacity as represented in the illustrative analysis, a possible mitigation is to reduce the effective number of independently tracked entities before execution. Future interfaces could enable explicit entity chunking, allowing users to group multiple objects (e.g., circling a cluster of entities that share a trajectory) so they can be articulated with a single, unified gesture. This acts as a technological proxy for working-memory chunking (Cowan, 2001): reducing the number of units that need to be individually tracked and expressed. This design hypothesis should be evaluated with higher-object-count events.

6.4.3. Separating Spatial Demonstration from Logical Annotation

One possible source of user fatigue is the cross-domain interference between visuospatial-motor execution (demonstrating motion) and verbal-analytical reasoning (specifying causal logic). Demanding simultaneous execution of

these qualitatively distinct sub-tasks may trigger the cognitive jamming described by P4. Future systems should decouple these phases to preserve cognitive bandwidth.

Upon detecting overload, the system could automatically transition to a two-step workflow. Users would first demonstrate all spatial trajectories purely via gesture, free from verbal interference. Subsequently, the system would prompt verbal annotations for causal logic (e.g., "What happens when the objects collide?"). For highly complex scenarios like *Event 4*, this workflow allows users to leverage the efficiency of parallel expression for spatial data while isolating the analytical demands of causal logic.

Ultimately, these design guidelines advocate for fluid, context-aware transitions between parallel and sequential expression *modes*, governed by the user's real-time cognitive capacity and coordination overhead.

7. Limitations and Future Work

While our work provides valuable insights into multimodal expression strategies for dynamic events in MR environments, several limitations should be acknowledged. First, our within-subjects study involved a relatively small and demographically concentrated university-based sample ($N = 24$; ages 21–31; 22 participants reporting Master's or doctoral education). The within-subjects design and Latin-square counterbalancing strengthen comparisons among conditions within this sample by reducing between-participant and order-related variability, but they do not compensate for limited population coverage. The sensitivity analysis in Appendix A.1 shows that the sample was adequate for within-participant effects of the size observed here ($d_z \approx 0.60$ with 80% power); detecting small effects would require a larger sample. Accordingly, the findings provide condition-specific evidence about relative performance and experience; they should not be interpreted as establishing a universal cognitive boundary of parallel expression across ages, educational backgrounds, or professional populations. Furthermore, the simplified dynamic events and controlled laboratory setting may not fully reflect real-world complexity. Future work should recruit larger and more diverse samples, including older adults and domain specialists, and evaluate these *modes* in ecologically valid contexts.

Second, our controlled quantitative evaluation was limited to events involving at most two objects ($N \leq 2$). The KLM-inspired analysis is therefore an illustrative abstraction of the observed two-object conditions, not

a validated model of scalability. Extending it to $N > 2$ requires assuming a regrouping-cost function and its parameter α; neither was measured or calibrated in this study. Moreover, KLM-style operators were originally developed for discrete, well-practiced actions, whereas our task involves continuous multimodal trajectories. Thus, the analysis can illustrate conditional relationships among strategies under its stated assumptions, but cannot estimate absolute completion times or an empirical crossover point. Future research should measure coordination costs directly for higher object counts and develop models tailored to continuous multimodal expression.

Third, all participants completed *events* in the same fixed order from simplest to most complex, while *mode* order was counterbalanced. Because each *event* level occurred at the same point in the session, changes across *events* may reflect growing task familiarity or fatigue rather than *event* demands alone. Pre-event practice sessions reduced these effects, but could not eliminate them entirely. The tests of *H2* and *H3* compare *modes* within the same *event*, so order effects that influence both *modes* equally cancel out. However, asymmetric practice or fatigue effects could still alter differences across *events*. For instance, the time savings were smaller at *Event 3* than at *Event 2* because *Sequential Multimodal* became faster, likely due to increased familiarity with its UI timeline. Furthermore, because *events* varied across multiple dimensions simultaneously (object count, timing, and causal interaction), we cannot isolate which specific factor drove the decline in subjective preference. Future studies should randomize *event* order and manipulate these variables independently.

Fourth, the conditions differed in interface support as well as in how they combined modalities: *Sequential Multimodal* offered information-block prompts and a timeline, and *Parallel Multimodal* offered error-free object instantiation through Wizard-of-Oz. This difference limits construct validity: the results reflect each interaction strategy as a whole, including its interface support. Future studies could vary the timing of modality use while holding interface support constant, for example by offering the same prompts and timeline in a parallel condition.

Finally, our evaluation prioritized user-centric subjective metrics, excluding objective system-side outcomes, such as the computational fidelity of the system's interpretation. Addressing this gap will require developing cross-modality metrics to quantify recognition/interpretation accuracy. Similarly, beyond verifying the coarse content of each expression, we did not evaluate how precisely participants expressed the events. The information-block

audit measures coverage, while completion time and word count measure efficiency; none of these metrics reflect accuracy. Therefore, faster expression does not imply clearer communication. Future work should reconstruct expressed events to quantify spatial and temporal errors, trajectory similarity, and event order accuracy. Additionally, we did not measure gaze- and hand-tracking accuracy in our specific hardware setup; future studies should include a metrological assessment of it (Salerno et al., 2026). Future systems should also explore dynamic adaptation strategies that adjust modality allocation in real time based on user performance and interpretation confidence.

Beyond these methodological extensions, future work should explore the proposed expression strategies in concrete applications: (1) in-situ XR authoring, where users combine speech, gaze, and gesture to create multi-object motion (Suzuki et al., 2020; Lee et al., 2025); (2) remote instruction and collaboration, where experts demonstrate how tools or components should move while explaining a procedure (Piumsomboon et al., 2019; Wang et al., 2020a); and (3) education and scientific explanation, where teachers and learners construct dynamic phenomena in MR (Victor, 2014; Radu, 2014). Together, these examples illustrate how our analysis framework can inform multimodal expression design across a broad range of real-world MR application scenarios, while experience gained from these real-world applications can, in turn, further refine and extend the theoretical framework.

8. Conclusion

In this work, we examined how users express dynamic events in Mixed Reality (MR), where rich spatiotemporal mental models must be externalized through limited output channels. We introduced a structured analysis that decomposes continuous dynamic events into five elemental information blocks and organizes them into a four-level complexity taxonomy, providing a foundation for designing and comparing multimodal expression strategies.

Using an abstract micro-world paradigm, we compared three complete interaction strategies (*sequential speech*, *sequential multimodal*, and *parallel multimodal* expression) across four events with increasing coordination demands. The results suggest a trade-off that depends on coordination demands. Under error-free speech–gaze fusion provided via a Wizard-of-Oz setup, *parallel multimodal* expression was on average faster and required fewer words across all tested events. However, its subjective advantages over *sequential multimodal* expression declined once events involved asynchronous

timing or causal interactions. For these events, we found no significant subjective differences between the two multimodal modes. This pattern suggests a redistribution of workload across modalities rather than an overall workload reduction. As coordination demands increased, the subjective benefits of parallel input diminished despite its persistent speed advantage.

We further examined this trade-off through a KLM-inspired sensitivity analysis, which serves only to generate hypotheses and illustrates how assumed coordination and regrouping costs could reduce the advantage of parallelism as object counts increase. Together, the empirical results and conditional analysis question the assumption that maximizing concurrent multimodal input always improves interaction. Future MR systems should therefore adaptively combine strategies, supporting parallel input for simple trajectories and structured sequential scaffolding for complex interactions. Testing the proposed scaling patterns with higher object counts remains important future work.

CRediT Authorship Contribution Statement

Aochen Jiao: Conceptualization, Methodology, Investigation, Formal analysis, Visualization, Writing – original draft, Writing – review & editing. **Yuxuan Li:** Investigation, Validation, Writing – original draft, Writing – review & editing. **Youpeng Zhang:** Methodology, Investigation, Writing – review & editing. **Hsin-Yi Wang:** Conceptualization, Methodology, Investigation. **Jiaxuan Wang:** Software. **Weitao Xu:** Supervision, Funding acquisition. **Eduardo Velloso:** Supervision, Writing – review & editing. **Nuwan Janaka:** Supervision, Methodology, Writing – review & editing. **Shengdong Zhao:** Supervision, Conceptualization, Methodology, Funding acquisition, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We gratefully acknowledge the late Hsin-Yi (Sam) Wang for his contributions to this work. His insights, dedication, and collaboration were invaluable

to the development of this research. We remember him with the deepest respect and gratitude, and dedicate this paper to his memory.

This research was supported by the CityU Start-up Grant (9610677).

Appendix A. User Study

Appendix A.1. Supplementary Robustness and Precision Checks

As a robustness check on the repeated-measures ANOVA, we fitted linear mixed-effects models to all subjective outcomes. Each model included fixed effects of *Mode*, *Event* , and *Event* × *Mode*, a participant-level random intercept, and uncorrelated participant-level random slopes for *Mode* and *Event* . We fitted the models by restricted maximum likelihood (REML) with sum-to-zero contrasts. For the Type III fixed-effect tests, we estimated denominator degrees of freedom with the Satterthwaite approximation and checked them with the Kenward–Roger approximation. Table A.5 reports the omnibus fixed-effect tests. These unadjusted p values are provided as a model-based sensitivity check; the family-wise Bonferroni results for the six *RTLX* components remain those reported in the main text. Nine of the 13 models were singular fits, meaning that at least one random-slope variance was estimated at zero. In practice, such a model behaves like one without that term, which is why the denominator degrees of freedom differ across outcomes. We used the same random-effects structure for every outcome and did not simplify it outcome by outcome. Across all 39 tests, the Satterthwaite and Kenward–Roger approximations gave the same decisions at $\alpha = .05$.

Because the six 7-point subjective measures were collected using single ordinal items, we also conducted Friedman tests comparing the three *modes* separately within each *event*. The complete omnibus results are shown in Table A.6. Where pairwise follow-ups are interpreted, the corresponding Wilcoxon signed-rank tests use the same within-event Bonferroni family of three design contrasts as the parametric post-hoc tests.

Table A.7 reports the estimated differences between the two multimodal *modes* in each *event* for the five outcomes used to evaluate *H2* and *H3*. The Event 4 estimates for the three subjective outcomes are close to zero, but the intervals quantify the remaining uncertainty; failure to reject a zero difference does not by itself establish statistical equivalence. We report ratios of condition means only for time and words, because the 7-point ratings have no meaningful zero point.

Table A.5: Linear mixed-effects robustness results. Entries are F (denominator degrees of freedom), p; numerator degrees of freedom are 2, 3, and 6 for *Mode*, *Event* , and *Event* × *Mode*, respectively. Denominator degrees of freedom were estimated with the Satterthwaite approximation; p values are unadjusted. †Singular fit (at least one random-slope variance estimated at zero).

Outcome	***Mode***	***Event***	***Event* × *Mode***
Preference†	7.78 (23.3), .003	2.16 (44.3), .107	3.89 (152.6), .001
Understanding†	9.01 (43.4), < .001	1.83 (91.9), .148	5.54 (207.0), < .001
Completeness	6.94 (23.5), .004	2.52 (32.5), .076	5.11 (135.8), < .001
Confidence†	6.45 (42.7), .004	2.24 (54.3), .094	3.73 (207.0), .002
Comfort†	6.91 (42.0), .003	3.57 (30.1), .026	2.24 (161.5), .042
Complexity	2.79 (23.7), .081	10.27 (34.9), < .001	2.72 (153.4), .015
Overall *RTLX*†	0.23 (22.3), .797	1.69 (36.0), .186	2.95 (148.9), .009
Mental	4.25 (23.6), .027	3.52 (32.0), .026	3.47 (138.6), .003
Physical	59.92 (24.5), < .001	1.51 (25.1), .235	2.02 (121.2), .068
Temporal†	2.04 (45.0), .142	0.70 (25.6), .563	1.02 (145.7), .415
Performance†	5.92 (22.9), .008	3.48 (66.3), .021	4.83 (177.4), < .001
Effort†	0.99 (42.9), .380	2.62 (35.6), .066	0.77 (174.8), .592
Frustration†	0.94 (23.3), .404	1.96 (33.9), .139	3.98 (152.5), < .001

The trend scores for the three subjective outcomes were not normally distributed (Shapiro–Wilk $p < .01$). We therefore repeated all five trend tests with two-sided Wilcoxon signed-rank tests on each participant's slope, using the same Bonferroni families. The corrected decisions did not change: Time, $W = 19.0$, $p_{bonf} < .001$; Words, $W = 44.0$, $p_{bonf} = .005$; Preference, $W = 43.0$, $p_{bonf} = .019$; Understanding, $W = 80.0$, $p_{bonf} = .135$; Completeness, $W = 27.5$, $p_{bonf} = .034$.

Finally, a paired-comparison sensitivity analysis shows that $N = 24$ provides 80% power to detect approximately $d_z = 0.60$ at a two-sided $\alpha = .05$, or $d_z = 0.70$ when $\alpha = .0167$ for a within-event Bonferroni family of three comparisons. This sensitivity result describes the detectable effect size under the achieved sample size; it is not presented as an a priori justification for participant recruitment. The trend contrasts are one-sample tests on 24 participant scores, so they have the same sensitivity. The *Perceived Understanding* trend ($d_z = -0.49$) is smaller than the effect we could detect with 80% power at the *H3* family level ($\alpha = .0167$). Its non-significance after

Table A.6: Non-parametric robustness checks for the six 7-point single-item measures. Entries are Friedman $\chi^2(2)$ (p).

Outcome	**Event 1**	**Event 2**	**Event 3**	**Event 4**
Preference	14.86 (< .001)	9.75 (.008)	2.59 (.274)	9.92 (.007)
Understanding	16.52 (< .001)	12.72 (.002)	0.74 (.692)	17.16 (< .001)
Completeness	7.19 (.027)	3.91 (.141)	11.70 (.003)	14.60 (< .001)
Confidence	11.36 (.003)	5.18 (.075)	4.20 (.122)	12.66 (.002)
Comfort	6.68 (.035)	10.55 (.005)	0.91 (.634)	9.24 (.010)
Complexity	4.52 (.104)	4.08 (.130)	3.03 (.220)	8.64 (.013)

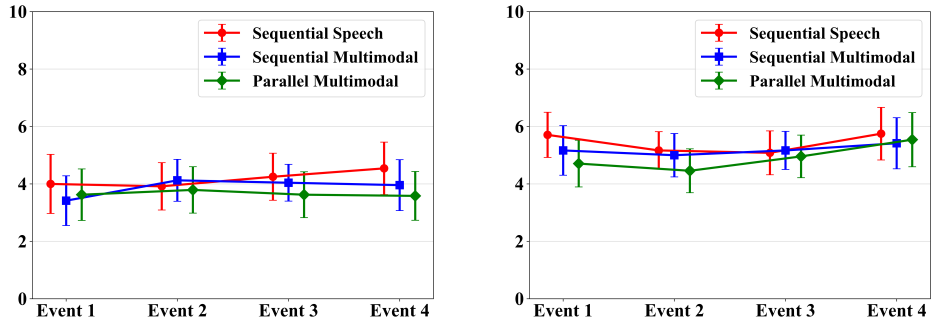
correction is therefore not evidence that there is no trend.

Table A.7: Per-event difference estimates for *Parallel Multimodal* (PM) minus *Sequential Multimodal* (SM) ($N = 24$). CIs are unadjusted. PM/SM is the ratio of condition means.

Outcome	**Event**	M_{PM}	M_{SM}	**PM − SM [95% CI]**	**PM/SM**	d_z
Time (s)	1	22.63	29.58	−6.95 [−13.65, −0.25]	.765	−0.44
	2	26.55	58.62	−32.07 [−40.16, −23.98]	.453	−1.67
	3	30.88	54.63	−23.75 [−33.17, −14.33]	.565	−1.06
	4	33.70	71.89	−38.19 [−48.76, −27.63]	.469	−1.53
Words	1	40.25	68.88	−28.63 [−55.75, −1.50]	.584	−0.45
	2	43.38	135.46	−92.08 [−121.85, −62.31]	.320	−1.31
	3	60.83	125.13	−64.29 [−93.37, −35.22]	.486	−0.93
	4	75.75	163.08	−87.33 [−122.81, −51.85]	.464	−1.04
Preference	1	6.125	4.750	1.375 [0.420, 2.330]	–	0.608
	2	5.417	4.792	0.625 [−0.066, 1.316]	–	0.382
	3	4.833	4.917	−0.083 [−0.580, 0.413]	–	−0.071
	4	5.125	5.333	−0.208 [−0.843, 0.426]	–	−0.139
Understanding	1	5.667	4.167	1.500 [0.466, 2.534]	–	0.612
	2	5.542	4.042	1.500 [0.877, 2.123]	–	1.017
	3	4.375	4.333	0.042 [−0.494, 0.577]	–	0.033
	4	4.875	4.833	0.042 [−0.789, 0.872]	–	0.021
Completeness	1	6.000	4.625	1.375 [0.479, 2.271]	–	0.648
	2	5.333	4.417	0.917 [0.110, 1.723]	–	0.480
	3	5.250	5.083	0.167 [−0.310, 0.644]	–	0.148
	4	4.875	4.917	−0.042 [−0.562, 0.479]	–	−0.034

Appendix A.2. Individual Indices of RTLX

In addition to the significant *RTLX* sub-dimensions reported in the main text, we also examined *Temporal Demand* and *Effort*. Neither dimension showed a significant main effect of *mode* or a significant *Event* × *Mode* interaction. Fig. A.10 visualizes these two non-significant sub-dimensions for completeness.



(a) *Temporal Demand*. (b) *Effort*.

Figure A.10: Non-significant *RTLX* sub-dimensions across events. *Temporal Demand* and *Effort* showed no significant main effects of *mode* or *Event* × *Mode* interactions. Error bars denote 95% confidence intervals.

Appendix B. KLM-Inspired Parameters and Illustrative Calculations

This appendix documents the illustrative operator values, event mapping, and step-by-step derivations used in the KLM-inspired sensitivity analysis in Section 6.3. The values are not fitted to the user-study data and do not constitute calibrated predictions of completion time.

Appendix B.1. Operator Calibration

We scale illustrative operators from the MR-validated measurements of Cabric et al. (Cabric et al., 2021), taking a speech chunk $S \approx 550$–600 ms as the base unit (1.0). Because the study task involves continuous trajectories rather than discrete actions, these values serve as relative abstractions rather than direct empirical calibrations:

- **Speech chunk** ($S = 1.0$)
- **Mental preparation** ($M = 0.5$)**:** Based on Simple Reaction ($M_s = 277$ ms; $M_s/S \approx 0.5$).
- **Gesture / motion** ($G = 0.8$)**:** Mean of Air Tap ($At = 427$ ms; $At/S \approx 0.7$) and Raise Hand ($Rh = 623$ ms; $Rh/S \approx 1.1$).
- **Gaze / head point** ($Z = 0.5$)**:** Discounted from Precise Head Point ($Php = 419$ ms; $Php/S \approx 0.7$) adjusting for lower required precision (Rayner, 1998).
- **System response** ($R = 0.3$)

Appendix B.2. Event-to-Model Mapping

Table B.8: Mapping from study event types to model parameters.

Level	N	c	c_{int}	g_{int}
Single Object	1	4	0	0
Two Object	2	4	0	0
Interaction	2	4	3	3

Appendix B.3. Speech-Only Expression

$$T_{\text{speech}} = M + (cN + c_{\text{int}})\, S = 0.5 + 4N + c_{\text{int}} \tag{B.1}$$

- **Single Object** (N=1, c_{int}=0): $0.5 + 4 = \mathbf{4.5}$
- **Two Object** (N=2, c_{int}=0): $0.5 + 8 = \mathbf{8.5}$
- **Interaction** (N=2, c_{int}=3): $0.5 + 8 + 3 = \mathbf{11.5}$

Appendix B.4. Sequential Multimodal Expression

$$T_{\text{seq}} = M + (cN + c_{\text{int}})\, S + (N + g_{\text{int}})\, G + (N + g_{\text{int}})\, R \tag{B.2}$$

- **Single Object** (N=1, c_{int}=0, g_{int}=0):
  $0.5 + 4.0 + 1(0.8) + 1(0.3) = \mathbf{5.6}$
- **Two Object** (N=2, c_{int}=0, g_{int}=0):
  $0.5 + 8.0 + 2(0.8) + 2(0.3) = \mathbf{10.7}$
- **Interaction** (N=2, c_{int}=3, g_{int}=3):
  $0.5 + 11.0 + 5(0.8) + 5(0.3) = 0.5 + 11.0 + 4.0 + 1.5 = \mathbf{17.0}$

Appendix B.5. Parallel Multimodal Expression

Using CPM-GOMS parallel-channel logic.

Appendix B.5.1. Single Object (N=1, no interaction)

- Step 1: $\max(S, Z) = \max(1.0, 0.5) = 1.0$
- Step 2: $\max(S, G) = \max(1.0, 0.8) = 1.0$
- End-point gesture ($G = 0.8$) + System response ($R = 0.3$)

$$T = 0.5 + 1.0 + 1.0 + 0.8 + 0.3 = \mathbf{3.6}$$

Appendix B.5.2. Two Object (N=2, no interaction)

- Step 1: $\max(2S, 2Z) = \max(2.0, 1.0) = 2.0$
- Step 2: $\max(2S, 2G) = \max(2.0, 1.6) = 2.0$
- End-point gesture ($G = 0.8$) + System response ($R = 0.3$)

$$T = 0.5 + 2.0 + 2.0 + 0.8 + 0.3 = \mathbf{5.6}$$

Appendix B.5.3. Interaction (N=2, c_{int}=3, g_{int}=3)

- **Object phase**: Step 1 ($\max(2S, 2Z) = 2.0$) + Step 2 ($\max(2S, 2G) = 2.0$)
- **Regrouping:** $M_r = 1.0$
- **Interaction phase**: Step 3 ($\max(2S_{\text{int}}, 2G_{\text{int}}) = \max(2.0, 1.6) = 2.0$)
- End-point gesture ($G = 0.8$) + System response ($R = 0.3$)

$$T = 0.5 + 2.0 + 2.0 + 1.0 + 2.0 + 0.8 + 0.3 = \mathbf{8.6}$$

Appendix B.6. Illustrative Chunked-Parallel Scenario ($N > 2$)

The following scenario uses Equation 4, where $T_{\text{r2}} = 4.0$ and $T_{\text{r1}} = 2.0$, and assumes a constant baseline regrouping cost $M_r = 1.0$. It is not derived from formal quantitative trials with $N > 2$:

- $N = 4$: $0.5 + 2(4.0) + 1(1.0) + 0.8 + 0.3 = \mathbf{10.6}$
- $N = 6$: $0.5 + 3(4.0) + 2(1.0) + 0.8 + 0.3 = \mathbf{15.6}$
- $N = 8$: $0.5 + 4(4.0) + 3(1.0) + 0.8 + 0.3 = \mathbf{20.6}$

Appendix B.7. Conditional Diminishing-Returns Derivation

For the conditional scenario in which the assumed regrouping function is $M_r(k) = t_{M_0} + f(k)$, where $K = \lceil N/2 \rceil$, the cumulative regrouping overhead is:

$$\Sigma_r(N) \;=\; \sum_{k=1}^{K-1} M_r(k) \;=\; (K{-}1)\, t_{M_0} \;+\; \sum_{k=1}^{K-1} f(k). \tag{B.3}$$

The first term is $O(N)$. If the assumed f is strictly increasing, the sum $\sum f(k)$ grows strictly super-linearly and the resulting mathematical scenario contains an intersection with the linear $T_{\text{speech}}(N)$. This conditional result does not establish that regrouping costs follow this function in actual use.

Derivation of an Illustrative Quadratic Scenario.. Substituting the illustrative function $f(k) = \alpha \cdot 2k$ (with $t_{M_0} = 0.5$ and $\alpha = 0.25$) yields $M_r(k) = 0.5 + 0.5k$. The cumulative cost for K rounds is:

$$\sum_{k=1}^{K-1} M_r(k) \;=\; (K{-}1)(0.5 + 0.25\,K) \tag{B.4}$$

Substituting $K = N/2$ (for even N) into Equation 4:

$$\begin{aligned} T_{\text{par}}(N) &= M + \tfrac{N}{2}\, T_{\text{r2}} + \big(\tfrac{N}{2} - 1\big)\big(0.5 + 0.25 \cdot \tfrac{N}{2}\big) + G + R \\ &= 0.5 + 2N + \big(\tfrac{N}{2} - 1\big)\big(0.5 + \tfrac{N}{8}\big) + 1.1 \\ &\approx 1.1 + 2.125\,N + 0.0625\,N^2 \end{aligned} \tag{B.5}$$

Calculated values for the illustrative quadratic scenario:

- $N = 4$: $0.5 + 8.0 + 1.0 + 1.1 = \mathbf{10.6}$
- $N = 6$: $0.5 + 12.0 + (1.0 + 1.5) + 1.1 = \mathbf{16.1}$
- $N = 8$: $0.5 + 16.0 + (1.0 + 1.5 + 2.0) + 1.1 = \mathbf{22.1}$
- $N = 10$: $0.5 + 20.0 + (1.0 + 1.5 + 2.0 + 2.5) + 1.1 = \mathbf{28.6}$

Appendix B.8. Sensitivity Analysis

Appendix B.8.1. Generalised Parallel Completion-Time Scenario

For the illustrative function $f(k) = \alpha \cdot 2k$, the regrouping cost becomes $M_r(k) = t_{M_0} + 2\alpha k$. For even N with $K = N/2$ bimanual rounds, the cumulative regrouping cost is $(K-1)(t_{M_0} + \alpha K)$. Substituting into Equation 4 gives the following scenario-dependent completion time:

$$T_{\text{par}}(N;\alpha) = 1.1 + \left(2.25 - \frac{\alpha}{2}\right) N + \frac{\alpha}{4} N^2 \tag{B.6}$$

Within this algebraic scenario, at $N = 2$ the quadratic and linear α-dependent terms cancel exactly ($-\alpha \cdot 2/2 + \alpha \cdot 4/4 = 0$), yielding an invariant relative ratio $T_{\text{par}}/T_{\text{seq}} = 5.6/10.7 = \mathbf{52.3\%}$. This is a property of the assumed model, not an empirical estimate.

Appendix B.8.2. Sensitivity to the Assumed α

Within this illustrative function, setting $T_{\text{par}}(N;\alpha) = T_{\text{speech}}(N)$ yields a scenario-specific crossover point N^*_{speech}:

$$\frac{\alpha}{4} N^2 - \left(1.75 + \frac{\alpha}{2}\right) N + 0.6 = 0 \tag{B.7}$$

Similarly, setting $T_{\text{par}} = T_{\text{seq}}$ yields the scenario-specific crossover point N^*_{seq}:

$$\frac{\alpha}{4} N^2 - \left(2.85 + \frac{\alpha}{2}\right) N + 0.6 = 0 \tag{B.8}$$

Appendix B.8.3. Sensitivity to Illustrative Micro-Level Operators ($N = 2$)

Within the illustrative operator scenario, we perturb each operator individually by $\pm 25\%$ from its baseline value. Because the parallel model at $N = 2$ relies on direct CPM-GOMS logic rather than a closed-form approximation, we recompute the `max` operators to account for bottleneck shifts.

- **Speech (S): 2.0** percentage point range. Decreasing S causes G to become the temporal bottleneck, partially negating the parallel compression advantage.
- **Mental (M): 1.1** percentage point range. Acts as a constant additive term; scaling it shifts the ratio slightly because the sequential denominator is larger.

- **Gaze (Z): 0.0** percentage point range. Z remains strictly smaller than S and does not trigger the bottleneck.
- **Gesture (G) and System (R): $\leq$ 0.2** percentage point range. Proportional changes affect both strategies similarly, preserving the ratio.

References

Alibali, M.W., 2005. Gesture in spatial cognition: Expressing, communicating, and thinking about spatial information. Spatial Cognition and Computation 5, 307–331.

Baddeley, A., 2020. Working memory. Memory , 71–111.

Bashar, M.R., Mutasim, A.K., Pfeuffer, K., Batmaz, A.U., 2026. Eyes on many: Evaluating gaze, hand, and voice for multi-object selection in extended reality, in: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–14.

Bolt, R.A., 1980. "Put-That-There": Voice and gesture at the graphics interface, in: Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques, pp. 262–270.

Brennan, S.E., Clark, H.H., 1996. Conceptual pacts and lexical choice in conversation. Journal of Experimental Psychology: Learning, Memory, and Cognition 22, 1482.

Cabric, F., Dubois, E., Serrano, M., 2021. A predictive performance model for immersive interactions in mixed reality, in: IEEE International Symposium on Mixed and Augmented Reality (ISMAR 2021).

Cai, R., Janaka, N., Kim, H., Chen, Y., Zhao, S., Huang, Y., Hsu, D., 2025. Aiget: Transforming everyday moments into hidden knowledge discovery with ai assistance on smart glasses, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, pp. 1–26.

Card, S.K., Moran, T.P., Newell, A., 1980. The keystroke-level model for user performance time with interactive systems. Communications of the ACM 23, 396–410.

Chen, J., Wu, X., Lan, T., Li, B., 2025. LLMER: Crafting interactive extended reality worlds with JSON data generated by large language models. IEEE Transactions on Visualization and Computer Graphics 31, 2715–2724.

Clark, H.H., 1996. Using language. Cambridge University Press.

Clark, H.H., 2003. Pointing and placing, in: Pointing. Psychology Press, pp. 251–276.

Clark, H.H., Brennan, S.E., 1991. Grounding in communication .

Cowan, N., 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. Behavioral and Brain Sciences 24, 87–114.

Cowan, N., 2010. The magical mystery four: How is working memory capacity limited, and why? Current Directions in Psychological Science 19, 51–57.

Dahlbäck, N., Jönsson, A., Ahrenberg, L., 1993. Wizard of Oz studies: Why and how, in: Proceedings of the 1st International Conference on Intelligent User Interfaces, pp. 193–200.

Davis, R.C., Colwell, B., Landay, J.A., 2008. K-Sketch: A "kinetic" sketch pad for novice animators, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 413–422.

De La Torre, F., Fang, C.M., Huang, H., Banburski-Fahey, A., Amores Fernandez, J., Lanier, J., 2024. LLMR: Real-time prompting of interactive worlds using large language models, in: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, pp. 1–22.

Du, R., Hersh, B., Li, D., Numan, N., Qian, X., Chen, Y., Zhou, Z., Chen, X., Ren, J., Bettridge, R.T., et al., 2026. Vibe coding XR: Accelerating AI + XR prototyping with XR Blocks and Gemini. arXiv preprint arXiv:2603.24591 .

Duchowski, A.T., 2017. Eye tracking methodology: Theory and practice. Springer.

Godden, E., Steedman, W., Pan, M.K., 2025. Robotic characterization of markerless hand-tracking on meta quest pro and quest 3 virtual reality headsets. IEEE Transactions on Visualization and Computer Graphics 31, 3025–3034.

Gray, W.D., John, B.E., Atwood, M.E., 1993. Project Ernestine: Validating a GOMS analysis for predicting and explaining real-world task performance. Human–Computer Interaction 8, 237–309.

Hart, S.G., 2006. NASA-task load index (NASA-TLX); 20 years later, in: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Sage publications Sage CA: Los Angeles, CA. pp. 904–908.

Hasson, U., Egidi, G., Marelli, M., Willems, R.M., 2018. Grounding the neurobiology of language in first principles: The necessity of non-language-centric explanations for language comprehension. Cognition 180, 135–157.

Heider, F., Simmel, M., 1944. An experimental study of apparent behavior. The American Journal of Psychology 57, 243–259.

Hincapié-Ramos, J.D., Guo, X., Moghadasian, P., Irani, P., 2014. Consumed endurance: A metric to quantify arm fatigue of mid-air interactions, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1063–1072.

Irawati, S., Green, S., Billinghurst, M., Duenser, A., Ko, H., 2006. "Move the couch where?": Developing an augmented reality multimodal interface, in: 2006 IEEE/ACM International Symposium on Mixed and Augmented Reality, IEEE. pp. 183–186.

Jackendoff, R., 2003. Précis of Foundations of Language: Brain, meaning, grammar, evolution. Behavioral and Brain Sciences 26, 651–665.

Jaimes, A., Sebe, N., 2007. Multimodal human–computer interaction: A survey. Computer Vision and Image Understanding 108, 116–134.

Jang, S., Stuerzlinger, W., Ambike, S., Ramani, K., 2017. Modeling cumulative arm fatigue in mid-air interaction based on perceived exertion and kinetics of arm motion, in: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 3328–3339.

John, B.E., Gray, W.D., 1995. CPM-GOMS: An analysis method for tasks with parallel activities, in: Conference Companion on Human Factors in Computing Systems, pp. 393–394.

Johnson-Laird, P., 1983. Mental models: Towards a cognitive science of language, inference, and consciousness. Harvard University Press.

Kaiser, E., Olwal, A., McGee, D., Benko, H., Corradini, A., Li, X., Cohen, P., Feiner, S., 2003. Mutual disambiguation of 3D multimodal interaction in augmented and virtual reality, in: Proceedings of the 5th International Conference on Multimodal Interfaces, pp. 12–19.

Kendon, A., 2004. Gesture: Visible action as utterance. Cambridge University Press.

Khan, A.A., Weidner, F., Rhee, J., Abdrabou, Y., Bianchi, A., Velloso, E., Gellersen, H., Newn, J., 2026. Gaze and speech in multimodal human-computer interaction: A scoping review, in: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems, pp. 1–22.

Kirsh, D., 2010. Thinking with external representations. AI & Society 25, 441–454.

Kirsh, D., Maglio, P., 1994. On distinguishing epistemic from pragmatic action. Cognitive Science 18, 513–549.

Kita, S., 2000. How representational gestures help speaking. Language and gesture 1, 162–185.

Kurby, C.A., Zacks, J.M., 2008. Segmentation in the perception and memory of events. Trends in Cognitive Sciences 12, 72–79.

Lee, J., Aleotti, F., Mazala, D., Garcia-Hernando, G., Vicente, S., Johnston, O.J., Kraus-Liang, I., Powierza, J., Shin, D., Froehlich, J.E., et al., 2025. ImaginateAR: AI-assisted in-situ authoring in augmented reality, in: Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology, pp. 1–21.

Lee, J., Wang, J., Brown, E., Chu, L., Rodriguez, S.S., Froehlich, J.E., 2024. GazePointAR: A context-aware multimodal voice assistant for pronoun disambiguation in wearable augmented reality, in: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, pp. 1–20.

Levelt, W.J., 1981. The speaker's linearization problem. Philosophical Transactions of the Royal Society of London. B, Biological Sciences 295, 305–315.

Levelt, W.J., 1993. Speaking: From intention to articulation. MIT Press.

Li, Y., Tag, B., Dai, S., Crowther, R., Dwyer, T., Irani, P., Ens, B., 2024. NICER: A new and improved consumed endurance and recovery metric to quantify muscle fatigue of mid-air interactions. ACM Transactions on Graphics 43, 1–14.

Lystbæk, M.N., Mikkelsen, T., Krisztandl, R., Gonzalez, E.J., Gonzalez-Franco, M., Gellersen, H., Pfeuffer, K., 2024. Hands-on, hands-off: Gaze-assisted bimanual 3D interaction, in: Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology, pp. 1–12.

McNeill, D., 1992. Hand and mind: What gestures reveal about thought. University of Chicago Press.

Nigay, L., Coutaz, J., 1993. A design space for multimodal systems: Concurrent processing and data fusion, in: Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems, pp. 172–178.

Oviatt, S., 1999. Ten myths of multimodal interaction. Communications of the ACM 42, 74–81.

Oviatt, S., Cohen, P., 2000. Perceptual user interfaces: Multimodal interfaces that process what comes naturally. Communications of the ACM 43, 45–53.

Oviatt, S., Coulston, R., Lunsford, R., 2004. When do we interact multimodally? Cognitive load and multimodal communication patterns, in: Proceedings of the 6th International Conference on Multimodal Interfaces, pp. 129–136.

Oviatt, S., Schuller, B., Cohen, P., Sonntag, D., Potamianos, G., 2017. The handbook of multimodal-multisensor interfaces, volume 1: Foundations, user modeling, and common modality combinations. Morgan & Claypool.

Pfeuffer, K., Mayer, B., Mardanbegi, D., Gellersen, H., 2017. Gaze + Pinch interaction in virtual reality, in: Proceedings of the 5th Symposium on Spatial User Interaction, pp. 99–108.

Pinker, S., 2003. The language instinct: How the mind creates language. Penguin UK.

Piumsomboon, T., Lee, G.A., Irlitti, A., Ens, B., Thomas, B.H., Billinghurst, M., 2019. On the shoulder of the giant: A multi-scale mixed reality collaboration with 360 video sharing and tangible interaction, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–17.

Posner, M.I., 1980. Orienting of attention. Quarterly Journal of Experimental Psychology 32, 3–25.

Radu, I., 2014. Augmented reality in education: A meta-review and cross-media analysis. Personal and Ubiquitous Computing 18, 1533–1543.

Rayner, K., 1998. Eye movements in reading and information processing: 20 years of research. Psychological Bulletin 124, 372.

Risko, E.F., Gilbert, S.J., 2016. Cognitive offloading. Trends in Cognitive Sciences 20, 676–688.

Rosenberg, K.T., Kazi, R.H., Wei, L.Y., Xia, H., Perlin, K., 2024. DrawTalking: Building interactive worlds by sketching and speaking, in: Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology, pp. 1–25.

Salerno, F., Ulrich, L., Maculotti, G., Moos, S., Genta, G., Vezzetti, E., Galetto, M., 2026. A metrological approach for augmented reality tooltip tracking assessment. Computers in Industry 175, 104430. doi:10.1016/j.compind.2025.104430.

Salvucci, D.D., Taatgen, N.A., 2008. Threaded cognition: An integrated theory of concurrent multitasking. Psychological Review 115, 101.

Santosa, S., Chevalier, F., Balakrishnan, R., Singh, K., 2013. Direct space-time trajectory control for visual media editing, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1149–1158.

Scaife, M., Rogers, Y., 1996. External cognition: How do graphical representations work? International Journal of Human-Computer Studies 45, 185–213.

Shi, J., Jain, R., Chi, S., Doh, H., Chi, H.g., Quinn, A.J., Ramani, K., 2025. CARING-AI: Towards authoring context-aware augmented reality instruction through generative artificial intelligence, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, pp. 1–23.

Suzuki, R., Kazi, R.H., Wei, L.Y., DiVerdi, S., Li, W., Leithinger, D., 2020. RealitySketch: Embedding responsive graphics and visualizations in AR through dynamic sketching, in: Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology, pp. 166–181.

Sweller, J., 1994. Cognitive load theory, learning difficulty, and instructional design. Learning and Instruction 4, 295–312.

Sweller, J., Van Merrienboer, J.J., Paas, F.G., 1998. Cognitive architecture and instructional design. Educational Psychology Review 10, 251–296.

Talmy, L., 2000. Toward a cognitive semantics: Concept structuring systems. volume 1. MIT Press.

Tse, E., Hancock, M., Greenberg, S., 2007. Speech-filtered bubble ray: Improving target acquisition on display walls, in: Proceedings of the 9th International Conference on Multimodal Interfaces, pp. 307–314.

Turkmen, R., Gelmez, Z.E., Batmaz, A.U., Stuerzlinger, W., Asente, P., Sarac, M., Pfeuffer, K., Barrera Machuca, M.D., 2024. EyeGuide & EyeConGuide: Gaze-based visual guides to improve 3D sketching systems, in: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, pp. 1–14.

Tütüncü, E.K., Zhou, Q., Brudy, F., Fitzmaurice, G., Anderson, F., 2026. PlayWrite: A multimodal system for AI supported narrative co-authoring through play in XR, in: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems, pp. 1–26.

Victor, B., 2014. Humane representation of thought: A trail map for the 21st century, in: Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology, pp. 699–699.

Wang, P., Bai, X., Billinghurst, M., Zhang, S., Han, D., Sun, M., Wang, Z., Lv, H., Han, S., 2020a. Haptic feedback helps me? A VR-SAR remote

collaborative system with tangible interaction. International Journal of Human–Computer Interaction 36, 1242–1257.

Wang, T., Qian, X., He, F., Hu, X., Cao, Y., Ramani, K., 2021. GesturAR: An authoring system for creating freehand interactive augmented reality applications, in: The 34th Annual ACM Symposium on User Interface Software and Technology, pp. 552–567.

Wang, T., Qian, X., He, F., Hu, X., Huo, K., Cao, Y., Ramani, K., 2020b. CAPturAR: An augmented reality tool for authoring human-involved context-aware applications, in: Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology, pp. 328–341.

Wang, Z., Shi, Y., Wang, Y., Yao, Y., Yan, K., Wang, Y., Ji, L., Xu, X., Yu, C., 2024. G-VOILA: Gaze-facilitated information querying in daily scenarios. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 8, 1–33.

Wei, S., Bloemers, D., Rovira, A., 2023. A preliminary study of the eye tracker in the Meta Quest Pro, in: Proceedings of the 2023 ACM International Conference on Interactive Media Experiences, ACM. pp. 216–221. doi:10.1145/3573381.3596467.

Weimer, D., Ganapathy, S.K., 1989. A synthetic visual environment with hand gesturing and voice input. ACM SIGCHI Bulletin 20, 235–240.

Wickens, C.D., 2002. Multiple resources and performance prediction. Theoretical Issues in Ergonomics Science 3, 159–177. doi:10.1080/14639220210123806.

Wickens, C.D., 2008. Multiple resources and mental workload. Human Factors 50, 449–455.

Zacks, J.M., Tversky, B., 2001. Event structure in perception and conception. Psychological Bulletin 127, 3.

Zhang, L., Pan, J., Gettig, J., Oney, S., Guo, A., 2024. VRCopilot: Authoring 3D layouts with generative AI models in VR, in: Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology, pp. 1–13.

Zhao, S., Tan, F., Fennedy, K., 2023. Heads-up computing moving beyond the device-centered paradigm. Communications of the ACM 66, 56–63.

Zwaan, R.A., Langston, M.C., Graesser, A.C., 1995. The construction of situation models in narrative comprehension: An event-indexing model. Psychological Science 6, 292–297.