

From Speech-to-Spatial: Grounding Utterances on A Live Shared View with Augmented Reality

Yoonsang Kim*

Divyansh Pradhan†

Devshree Jadeja‡

Arie E. Kaufman§

Center for Visual Computing, Stony Brook University

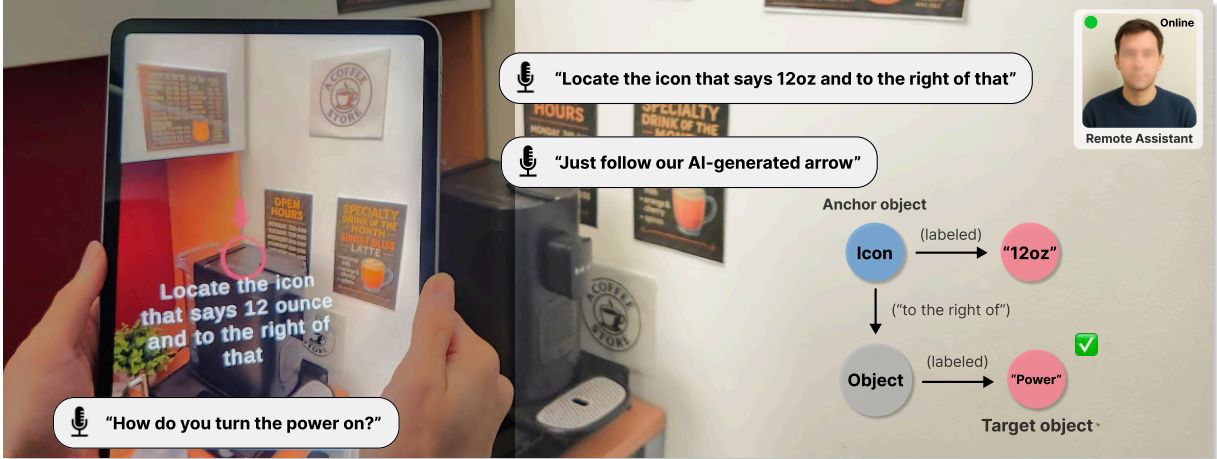


Figure 1: Concept illustration of Speech-to-Spatial, disambiguating verbal descriptions of a referent and situating AR visual guiders leveraging object-centric graph, and LLM-based reasoning in a remote assistance scenario.

ABSTRACT

We introduce Speech-to-Spatial, a referent disambiguation framework that converts verbal remote-assistance instructions into spatially grounded AR guidance. Unlike prior systems that rely on additional cues (e.g., gesture, gaze) or manual expert annotations, Speech-to-Spatial infers the intended target solely from spoken references (speech input). Motivated by our formative study of speech referencing patterns, we characterize recurring ways people specify targets (Direct Attribute, Relational, Remembrance, and Chained) and ground them to our object-centric relational graph. Given an utterance, referent cues are parsed and rendered as persistent in-situ AR visual guidance, reducing iterative micro-guidance (“a bit more to the right”, “now, stop.”) during remote guidance. We demonstrate the use cases of our system with remote guided assistance and intent disambiguation scenarios. Our evaluation shows that Speech-to-Spatial improves task efficiency, reduces cognitive load, and enhances usability compared to a conventional voice-only baseline, transforming disembodied verbal instruction into visually explainable, actionable guidance on a live shared view.

Index Terms: Speech, Spatial Interface, Remote Collaboration, Spatial Referencing, Augmented Reality, Large Language Models.

1 INTRODUCTION

Remote assistance often relies on spoken instructions [14], which are ambiguous when designating referents. Prior works show that language alone is insufficient, and phrases such as “this one” or

“over there” remain under-specified unless paired with embodied cues such as gesture or gaze [1, 2, 27, 40]. Recent studies demonstrate that combining verbal and visual cues, aided by LLMs, enhances the clarity of verbal descriptions [28, 41, 64]. Beyond gesture, relational context also grounds meaning: ontology-driven approaches model multi-dimensional attributes such as space, time, activity, actor, and object [34], and relational priors and scene graphs yield dense, spatially consistent reconstructions [53]. These works suggest that ambiguity can be reduced through structural analysis of how entities relate both spatially and semantically.

While graphs and multimodal cues clarify communication, eXtended Reality (XR) collaboration introduces practical challenges. Visualizing embodied gestures or remote users’ avatars requires specialized hardware such as motion tracking and immersive Head-Mounted Displays (HMDs). These requirements slow the adoption of immersive solutions in remote technical support, where practitioners often fall back to 2D displays, hand-held device-based communication, or manual cursor annotations. Other domains employ chatbot-style AI with no direct human-to-human assistance or embodied presence. These constraints underline the need for lightweight approaches that can operate in speech-only conditions.

We propose **Speech-to-Spatial** to address these challenges. Speech-to-Spatial is a framework that grounds verbal-only instructions into Augmented Reality (AR) visual guidance (Fig. 1). Building on the existing literature on the ambiguity of spatial expressions and our formative study, we categorize four recurring referential types – Direct Feature-based, Relational, Memory-based, and Chained. Speech-to-Spatial reconstructs a relational graph, maps spoken references, and auto-generates AR indicators that disambiguate spatial descriptions. This transforms transient utterances into persistent, spatially contextualized guidance—without requiring additional gear, cues, or manual human annotations.

We showcase the use cases of Speech-to-Spatial across three collaborative scenarios: Remote maintenance instructions, Indoor navigation, and AI personal assistance. Our evaluation shows measur-

*e-mail:yoonsakim@cs.stonybrook.edu

†e-mail:divyansh.pradhan@stonybrook.edu

‡e-mail:devshreehardik.jadeja@stonybrook.edu

§e-mail:ari@cs.stonybrook.edu

able gains in task efficiency, accuracy, reduced cognitive load, and heightened usability for Speech-to-Spatial, compared to the voice-only assistance baseline. Our contributions are as follows:

- **An end-to-end pipeline for speech disambiguation** We elucidate the intent behind speech-only instructions using graph-traversal reasoning and provide situated visual guidance through an AR indicator.
- **Spatial description patterns in remote instructions** We derive five recurring language patterns in spatial descriptions (Direct-feature, Relational, Memory, Chained, and Deictic), stemmed from established studies of Language.
- **Demonstration on collaborative scenarios** We showcase the applications of Speech-to-Spatial-Maintenance, Navigation, and Personal assistance, and its potential.
- **Empirical evaluation** We assess and compare the performance of Speech-to-Spatial-cognitive load and usability over an existing remote assistance baseline (voice-only), with implications for integration into existing methods.

By unraveling the ambiguities of verbal interaction in remote assistance and automatically enriching the channel of communication by providing additional visual guidance, Speech-to-Spatial advances remote assistance beyond disembodied voice interfaces toward spatially grounded and visually explainable experiences.

2 RELATED WORK

2.1 Remote Assistance and Information Sharing

Remote assistance for procedural tasks has relied on phone or video calls with verbal instructions. While videos provide basic visibility, they lack shared spatial grounding, making it hard to identify precise referents or actions [14]. AR and MR research have introduced in-view overlays and annotations to reduce miscommunication [23]. Embedding step-by-step instructions into AR views improved correctness in maintenance tasks [45, 54]. Surveys further confirm that MR-based remote assistance reduces task errors and increases efficiency [17]. In remote AR guidance, augmenting visual content (e.g., avatar, view-point ray, gestures, pointers) onto the physical realm has shown to facilitate information communication and collaboration [18, 20, 33, 50, 59, 69]. Another set of work shows the use of gestures to enrich data communication—for presentations [4] as well. Yet, many commercial platforms remain dominated by 2D desktop or mobile interfaces, relying on manual pointers or cursors even with the use of AR, and human-authored annotations for grounding [14, 63, 68].

This motivates Speech-to-Spatial to bridge the gap between manual-annotated AR remote assistive technology and conventional speech-only remote guidance. We aim to automatically disambiguate spatial referencing in verbal descriptions using an LLM and a graph-structured reasoning pipeline, enriching a single-channel (verbal) remote guidance to a dual-channel (visual, verbal).

2.2 Spatial Referencing in Spoken Language

Prior works have established that spatial language contains statements that refer to objects with respect to a reference frame [42, 43, 62], and have shown that people adopt different perspectives and frames of reference—viewer-centered, object-centered, or environment-centered—and that mismatches can cause misunderstandings [5, 30, 57]. Taylor and Tversky indicate that spatial descriptions are inherently perspective-dependent [66]. In wayfinding scenarios, where only the objects near the user are visible, the viewer and the environment-centered (Cardinal directions—North, East, West, South) expressions were used. In contrast, when a scene can be perceived within a single viewpoint, object-centered spatial referencing was more prevalent (e.g., “Object A to the right of Object B”). Recent interactive and AI systems adopt the object-centered spatial expressions, by pairing language with visual cues

to indicate target referents [9, 11, 26, 31]. Target (the referred object) and Anchor (the figural object used to refer to the Target) attributes (e.g., color, size, shape), relational language, and surrounding visual context further support disambiguation of referents [7, 13, 28, 58]. These intriguing language patterns of spatial descriptions, and their use cases suggest that explicit spatial anchoring and relational structure play a role in resolving under-specified spoken references.

We motivate our object-centered graph representation from these human spatial description patterns, and use a multimodal LLM to interpret ambiguous target references during remote assistance. As a visual guidance system that situates instructions onto a referred target, we extend a verbally-instructed guidance to visual communication, enabling robust grounding and disambiguation.

2.3 Multimodal Cues and Disambiguation

The seminal work, “Put-That-There” [2], established how speech and gesture interrelate each other and treat deictic terms as temporary variables grounded by pointing to spatial targets. Recent systems extend this principle to visualization and XR domains by fusing embodied cues. Han and Issac leverage deictic references (e.g., this, that, here, there) to enrich interaction for visual analytics [24]. GazePointAR employs a gaze and gesture-aware personal assistant to disambiguate under-specified spoken queries in real-time [40]. GesPrompt uses the synchronization between the temporal dimension, speech, and co-speech gestures, to capture richer spatial-temporal intent [27]. Bovo et al. revisits the “Put-That-There,” paradigm for XR information placement with scene semantics and head and pointing cues to interpret under-specified commands [3].

Recent works leverage LLM agents to bind multimodal context (visual, audio, gesture, interaction history) to reason over spatial context [24, 40, 41, 47, 64, 70], and automate grounding. While the fusion of context can mitigate the ambiguity, such techniques assume reliable tracking of gaze, gesture, avatars, or increased sensing computations. Rather than relying on embodied sensing with additional cues, Speech-to-Spatial treats speech as the sole instructor channel (lightweight) and resolves ambiguity through a structured referent reasoning pipeline, then generates visual overlays that approximate the disambiguating role of embodied cues.

2.4 Intelligent Grounding in XR

Recent work integrates scene understanding and language models to automate grounding. Guided Reality demonstrates how LLMs and vision models generate visually-enriched task cues embedded into the scene [70]. Complementary works on dialogue augmentation [8] and XR-Objects [12], explore embedding relational semantics from conversation into situated overlays. In parallel, relational scene representations including scene graphs, are increasingly leveraged to capture multi-object relationships and support grounding beyond single entities [9, 15, 32, 38, 53, 55]. These 3D graphs provide hierarchical structure, enabling more transparent relationship predictions than neural representations. ConceptGraphs utilize open-vocabulary detector and scene graphs, enabling systems to query using relational prompts or identify targets [22].

These threads demonstrate the use of multimodal LLM-driven context-aware reasoning. Speech-to-Spatial builds on this trajectory, but focuses on the intricacy of verbal descriptions in remote instructions. We ground a speech signal into lightweight, AR anchored visualizations, without requiring additional modalities.

2.5 Memory, Recall, and Situatedness

Visual grounding extends beyond task performance into the domains of long-term memory and recall. Research indicates that pairing speech with visual cues improves information retention and reduces errors. Specifically, Lukianova et al. found that images paired with text in AR instructions, significantly boost recall over

text-only conditions. Visual information is processed by the brain more efficiently than linguistic tokens due to their natural semantics over learned symbols [48]. Situated visualization studies also show how context-bound representations support recollection and decision-making [6, 21, 36]. In AR, spatial markers further support task switching and resumption by visually situating attention and helping users return to a spatial context even after task interruption [49]. Memory-oriented systems explore how interaction histories can be captured and reused [56, 61]. Memoro memorizes prior dialogue to support verbal remembrance nudging [72]. OmniQuery extends this idea by connecting visual memory with other contextual cues for information search and retrieval [46].

These works suggest that situated persistent memory traces can be used for longer-term usability and spatial reasoning. Speech-to-Spatial builds on this by treating every referent as incremental semantic memory. That is, every action and relationship between a user and a referent is stored as a memory. By maintaining an object-centric interaction history within a 3D knowledge graph, Speech-to-Spatial can trace prior instructions and context, enabling effective disambiguation of under-specified spatial descriptions.

3 DESIGN OF SPEECH-TO-SPATIAL FRAMEWORK

3.1 The Need for Referent Disambiguation in Speech

The core challenge in remote assistance is communication clarity. The meaning of an expert’s instructions must be quickly understood and acted upon for a collaboration to be successful. However, conventional remote support solutions rely on an asymmetric setup where a local worker streams a visual feed while a remote expert provides speech-only guidance. Since spatial language is inherently ambiguous, this triggers repeated back-and-forth to clarify the intended target and what action should be taken. Recent approaches (e.g., video calls with marking tools or AR tele-assistance) partially address this by enabling experts to add visual annotations, but these cues are typically created manually, adding extra burden.

Speech-to-Spatial aims to address this communication bottleneck by designing a disambiguation layer that clarifies the referent behind the remote expert’s spoken instruction. We aim to *remove the burden of manual visual annotations* by introducing an automated *disambiguation* pipeline based on speech, and transforming the expert’s *speech* into spatially grounded visual indicators.

3.1.1 Understanding the Language Pattern in Speech Guided Remote Assistance

We ground our analysis in established perspectives on spatial reference frames drawn from prior literature: viewer-centered, object-centered, and environment-centered [5, 30, 42, 43, 57, 62, 66]. To understand how these perspectives manifest in remote assistance, we examine the recurring linguistic strategies and communication patterns that arise when an expert guides a worker through verbal instructions. We conduct a preliminary formative study of remote verbal instructions with a shared screen view to mimic a remote assistance setting. The insights derived from the study will inform the design of Speech-to-Spatial, a framework that disambiguates the spoken spatial references of target objects in remote instructions and produces visual+speech-supported guidance.

The core aim of the formative study was to analyze the emerging spatial description patterns during remote instructions as the first step, and to base this insight on the design of Speech-to-Spatial, and evaluate the effectiveness of our combined approach with AR.

3.1.2 Study Setup and Procedure

We recruited 9 participants (academic researchers, engineers, and students; 8 male, 1 female; aged 27-34; P1-9), paired as instructor (giving directions) and follower (executing them). P1 volunteered to be the designated follower across all sessions (P2-9 being instructors). A 15-minute session with 30 instructions was conducted remotely via Zoom [71] with 2D screen sharing as the shared

workspace. The follower was instructed to follow commands literally (dull following), without interpretation, and respond only minimally to the instructor for confirmation. This was to observe and classify spoken patterns of each single turn (a single-trip: query-and-response) conversation, as a multi-turn conversation is a composite that involves more than one single-turn conversations with prior conversation context. The first author (of this research) observed silently, recording transcripts and notable referring expressions. After each session, participants were interviewed about their strategies and difficulties based on the author’s session notes. The notes were categorized into high-level themes (Sec. 3.1.3).

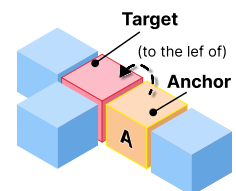
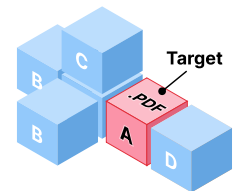
The tasks were designed as an instruction-following activity on a shared desktop view. The 2D screen of the follower was shared with the instructor, representing a remote assistance scenario. For each trial, the instructor was privately informed of the randomly chosen target, ranging from an empty folder to an existing file/icon on the Desktop. The instructor, then, guided the follower to select the item by moving their mouse cursor. To avoid inevitable ambiguity from the non-distinctive feature of a target, the follower was asked, before the study, to create a set of empty folders on their desktop, each named with a unique single letter in the English alphabet. This setup ensured that the instructor always had the freedom to choose any referencing method to refer to a target, without indirectly converging on using a specific referencing method. The first 15 tasks were performed (1) without access to the annotation tool, enforcing speech-only communication, and the other 15 were performed (2) with annotation enabled to examine any shift in participants’ strategies upon access to the annotation tool.

3.1.3 Findings and Implications

The study revealed four outstanding patterns of spatial reference, which we thematically coded as Direct Feature, Relational, Memory, and Chained references, following similar groupings of spatial linguistic expressions: Using figural/landmark objects to describe a scene [52, 65, 66], describing target features [52], using relational/relative descriptions [19, 44], referring back to previously interacted targets [10, 42, 66], and deictic references [43]. We also observe the use of deictic referencing (e.g., “that”, “it”) when the drawing/annotation tool was enabled. We report the total occurrence of each pattern across all tasks, not only the initial descriptions (e.g., “click the pdf file”), but including the recovery attempts (e.g., “No, one to the right of it”). One task may involve one or more patterns. The patterns are based on our observation notes (N=187).

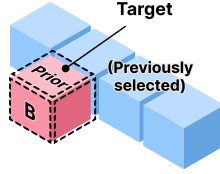
Direct Feature. The instructor attempted to describe the target directly (57.6%) through its intrinsic attributes and features in their initial trials. This includes references to distinct attributes such as the color (“the *red* file”), type (“the *PDF* file”), or labels (“the file named *A*”). Direct Features were effective when the distinguishing attribute was conspicuous. However, when multiple items shared similar features (e.g., similar-colored folders), instructors added additional descriptions, or shifted to other referencing techniques, suggesting that feature-based references provide an important baseline, yet, are fragile for target reference on its own.

Relational. When Direct Feature referencing was insufficient, the participants used this type (31.2%). It uses an anchor referent/object that is more conspicuous and acts as a landmark to the target. Phrases such as “the *one to the left of the yellow file*.” Relational referencing allowed instructors to dis-

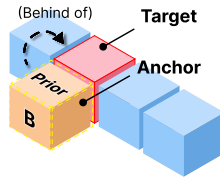


ambiguate targets even when features overlapped, but it required the follower to correctly identify the anchor object first. When the follower still failed to identify the anchor, the instructor resorted back to guiding the follower’s mouse cursor as the anchor point. For example, “*move a little more to the left*” or “*to your right*” (micro-guidance). The prevalence of this pattern indicates the importance of modeling spatial relations.

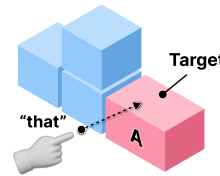
Memory. Participants used prior knowledge to indicate an anchor referent (a distinct referent to indicate the target) (11.2%). This pattern refers back to the objects that had been mentioned or manipulated earlier in the session, the short-term shared knowledge between the instructor and the follower. An example phrase could be: “*the file we previously selected*”. This strategy highlights how the anchor point in referring a target object is not limited to an object’s distinct attributes, but can also be based on the shared experience the two collaborators engaged in. This shared experience then becomes a resource for grounding cues as well. However, one of the participants indicated that the Memory pattern caused them additional cognitive overhead, as they tried to recall their interaction history. This referencing motivates our system design to retain interaction history as part of the grounding step.



Chained. This is a composite pattern. In a single utterance, it layers Direct Feature, Relational, and/or Memory-based cues. For instance: “*the folder behind the one we selected earlier*” (Relational, Memory). These references emerged when the description complexity of the anchor or target increased, or when initial guidance attempts failed. While effective, an elongated chain of instructions was reported to frustrate both the follower and the instructor, often leading them to merely follow mouse cursor-based micro-guidance. The Chained referencing pattern illustrates the need for a compositional representation such as a graph that can integrate cues beyond spatial relations between referents. Its statistics are broken down into its individual patterns.



Deictic. This pattern was the most prominently used with the visual annotation tool. For example, “*that one*”, “*it*” (while visually marking the target). However, this pattern is dependent on additional cue such as visual indication or explicit gesture pointing [2, 27, 40]. As we disambiguate solely on speech cue, we consider this pattern as out of our scope, and discuss in our future work discussion (Sec. 6).



3.2 Design Rationale

Our formative study validated that the linguistic spatial patterns established in (Sec. 2.2) also emerge in remote verbal guidance scenarios. These descriptions relied on (1) visual features or attributes of objects, (2) relations among entities, and (3) shared interaction context. We use a graph representation that retains these patterns and supports structured interpretation and reasoning over referred targets in spoken instructions. Leveraging this representation, we *disambiguate the referred target in speech instructions* of remote assistance and project AR guidance onto the resolved referent for *explicit visual guidance*. We detail the design considerations for our framework in the following:

D1 Spatial representation for referent resolution. A referent-centered 3D spatial representation that can retain its own attributes

(e.g., 6DoF transformation, color, size, shape), and provide a way to connect and traverse other neighboring referents. The structure must be able to represent the three spatial expressions (Direct, Relational, and Chained).

D2 Retain prior interaction context. The spatial representation not only retains what attributes an entity/referent has, but also the temporal grounding (interaction history–Memory). The system must be able to retrieve the prior activities applied to each referent in a form of an episodic memory.

D3 End-to-end generation of visual indicator. To address the adoption barrier of manual annotation, the system must convert referential language into a concrete spatial output automatically. This requires an end-to-end pipeline that maps utterances to candidate referents and emits an explicit visual indication, reducing reliance on repeated verbal clarification.

D4 Lightweight instrumentation. To remain practical across diverse AR deployment settings, the referent disambiguation from a verbal instruction, must not require any additional embodied cues (e.g., gaze, gesture) or hardware-specific dependencies, while capturing comprehensive context about the user (e.g., space, time, activity, intent, referent) [29, 35, 39].

In designing Speech-to-Spatial, we treat speech as an expressive specification of referential intent, and convert it into automatic anchored visual guidance to make spatial grounding attainable in lightweight settings without additional cues such as gaze or gestures. We aim to disambiguate a target from verbal expressions, into visually grounded guidance using graph-based reasoning with an LLM and AR, enriching remote communication beyond manual visual annotations with verbal instructions.

3.3 System Implementation

System Overview. Speech-to-Spatial follows an end-to-end workflow: (1) spoken instructions are transcribed and interpreted; (2) the target, anchor (if applicable), and the descriptions of each object are parsed; (3) an object-centric graph is constructed from the visible objects and referents (D1); (4) attributes per node are assigned from the visible features; (5) the referred object is assigned with its requested interaction history (D2); and (6) an augmented visual overlay is situated onto the physical environment (D3), without necessitating other cues beyond speech (D4).

At the core of Speech-to-Spatial lies the object-centric graph. It allows every visible physical object to retain multi-dimensional attributes–space, time, action, intent, and actor–capturing the history of interactions, while also encoding its own feature descriptors (e.g., color, shape), and spatial relations between neighboring objects. This persistent object-based representation supports the integration of the four referencing patterns and facilitates the disambiguation of object references during remote verbal communication. The pipeline is illustrated in Fig. 2. Please refer to our Supplementary Material for more implementation details.

Implementation Environment. We offload the computation overhead from the client’s device (via a client-server architecture). On the client’s device, Unity AR Foundation was used, and we use a custom Python server for the reasoning backend. Unity handles the visual capture, voice recording, 6DoF pose extraction of objects, and AR anchoring of contents. Objects are represented internally as JSON nodes before being incorporated into the object-centric graph on the server. Speech is transcribed using Whisper [16], and utterance parsing, reasoning, and resolution are handled by GPT-4.1, followed by the text-embedding-3-small model, which generates the embeddings for object attributes. Gemini 2.5-Flash was used to localize, segment, and classify objects, and perform visual analysis of the scene. Embeddings of object attributes are cached and reused for optimality.

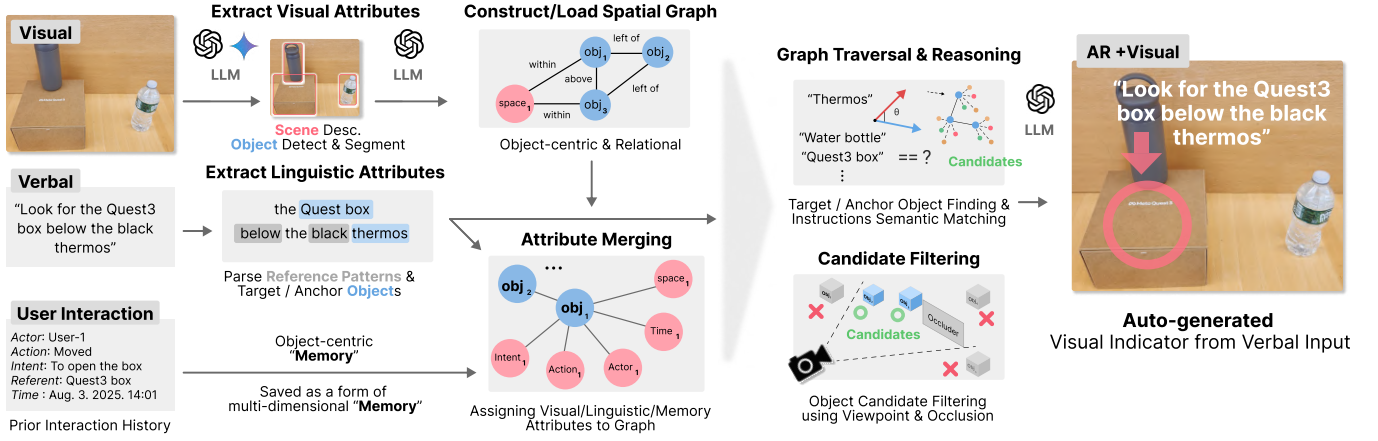


Figure 2: End-to-end pipeline of Speech-to-Spatial: From speech with visual inputs and prior memories (if present), Speech-to-Spatial extracts the visible, linguistic attributes, associates them with a relational graph representation, and generates an AR visual indicator.

3.3.1 Remote Instruction Parsing and Attribute Extraction

When Speech-to-Spatial receives a spoken instruction, it parses the utterance into a pre-defined structure with an LLM. The structure encodes the linguistic dimensions: $\{targetobject\}$, one or more $\{anchorobject\}$, their respective object $\{class/label\}$, $\{description/features\}$, a $\{relationalphrase\}$, and any $\{action\}$, $\{intent\}$, or $\{temporal\}$ cues. As shown in Fig. 3, “The black thermos above the Quest 3 box, next to the Poland spring water we talked about a minute ago” yields the target node (“thermos”), the anchor node (“box”), associated with the attributes of each node, and their Direct Feature, Relational, and Memory-based relationship. Generic nouns such as “thing”, or “it” are not treated as a target label or description, but rather as the question of the referent.

3.3.2 Object-centric Relational Graph Construction

Upon the first encounter with a scene and the localization of an object, we build a relational graph, mapping the logical node-object to the physical object (Fig. 4). The graph is retained across AR sessions enabling users to permanently refer back to their prior interactions with the objects. Speech-to-Spatial bypasses the graph construction step once it already holds a representation of the scene.

Object Detection and Segmentation. We localize the visible objects in the captured snapshot via a vision-capable LLM. It is prompted to localize all visible objects, as well as to extract their descriptive features (e.g., color, class-label, shape). Specifically, we guide the LLM to localize the parsed target and anchor objects.

Graph Construction. Once objects in a scene are localized, and their features are extracted, a relational graph is constructed. Every localized physical object is registered as a logical node in the graph, and the nodes are connected by spatial relationships between their neighboring nodes. The spatial graph is constructed using the six spatial relational properties: “left”, “right”, “above”, “below”, “in-front-of”, and “behind-of”. We define an object to be “in relation” to another only when it is within a half-meter radius ($r=50cm$). The 3D center point of a localized node, derived from the axis-aligned 3D bounding box of a target object, is used to determine the spatial relation between objects, for simplicity.

Assigning Attributes to Graph Nodes. We assign identified features (e.g., color, class-label) of each object extracted from earlier steps, to logical node attributes and its spatial context (the description of the overall scene the objects are in; e.g., “Desk with laptop and coffee on the side.”). Then, we initialize each graph node with an empty memory field (Fig. 3), which binds the temporal footprint and action/intent history of user(s) when an interaction occurs.

The ontology-driven structures [15, 22, 34, 53] have shown that graphs can preserve explicit relations and histories, ensuring that references are resolved through paths that remain interpretable and

explainable, unlike flat vector-based representations. By combining the concept of a scene graph with multi-context—space, time, actor, action, intent [29, 35, 60]—we represent spatial relationships across objects as well as their interaction histories and context.

3.3.3 Referent Inference and Reasoning

Compositional and Chained Reasoning. Speech-to-Spatial resolves utterances that involve multiple dimensions of reference. Each node in the object-centric graph consists of not only what the object is, but also where it is, what has been done to it, and when; on top of the spatial relationship between its neighboring objects. With the integration of relational, spatial, temporal, action, and intent-based dimensions, the graph supports compositional reasoning across multiple anchors and chained references. For instance, “the cube behind the sphere and in front of the machine” can be resolved by intersecting relational paths from the two anchor nodes, while “the bolt next to the panel we fixed earlier” requires traversing both spatial and temporal-action histories.

Semantic Embeddings and Attribute Matching. Relational properties are fixed to six terms (e.g., two of which are “left” and “right”). However, Direct Feature (‘label’, ‘description’ in Fig. 3) or Memory-based referencing rely on descriptive language, not constrained to a set of pre-defined vocabularies. To handle this variability, Speech-to-Spatial computes semantic similarity between each graph node and the parsed attributes of the target/anchors (cosine similarity between vectors). We select the top five candidate nodes ($k=5$) with highest attribute description similarity, while satisfying relational and memory attribute alignment. Then, we perform an LLM-based reasoning to pinpoint the referred node. The first-pass semantic similarity node filtering, not only reduces the input context counts passed to the LLM in the final step, but also provides robustness to linguistic variation in referents beyond Bag-of-Words or naive keyword matching.

Viewpoint-aware Candidate Filtering. Speech-to-Spatial employs object-level frustum and occlusion-culling of object nodes, to minimize the similarity computation checks of candidate nodes in the graph. Similar to the culling techniques in graphics engines, only the nodes that are in view are retained, while those outside are discarded. Then, naive occlusion culling is performed by casting an AR ray (physical surface depth checks) to the known object-node position, for depth-testing following the pseudo-code:

if ($D_A > D_B + \Delta_{scale}$) then: *occluded*
 (where $A = cam.to_obj_dist$, $B = ray.hit_dist$,
 $\Delta_{scale} = target.obj_scale$)

At the end of the reasoning pipeline, we employ a fallback mechanism in the case of conflicting target referencing or reasoning failures. We leverage an evaluation LLM agent to verify whether the

The black thermos above the Quest3 box, next to the Poland spring water we talked about a minute ago

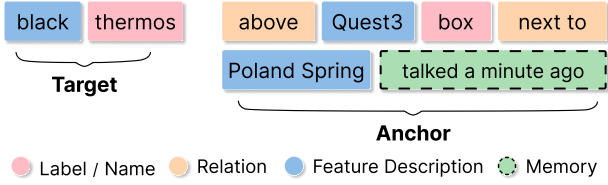


Figure 3: Attribute parsing: Transcribed text of verbal instructions is extracted into a structure via LLM.

final candidate referent satisfies all the conditions without ambiguities. Upon conflict, it falls back to the raw transcription anchoring, instead of a visual pointer to avoid any uninformed guidance.

3.3.4 Capturing Interaction History

Retaining Action Footprints. A central principle of Speech-to-Spatial is that memory is not an auxiliary log but an integral part of the graph itself. Each node accumulates a persistent history of interactions: the actions taken (e.g., “moved”, “rotated”), the actors involved, and the temporal footprint of those actions. By storing these directly onto object-nodes, Speech-to-Spatial ensures that references such as “the panel we looked at yesterday” or “the folder we opened earlier” can be naturally resolved by traversing temporal-action attributes. This approach turns every object into a site of accumulated memory, anchoring its evolving state across time.

Revisability of Object-nodes. The object nodes of the graph are preserved across sessions. However, the attribute of an object can be updated through a user interaction. When an object is transformed, moved, or altered, Speech-to-Spatial updates its attributes in place. This interaction ‘Action’ is recorded in the interaction history of the object-node, and is appended to its existing interaction data. This supports the Memory-based referencing pattern, enabling the recall of a referent based on prior interactions.

3.3.5 Anchoring and Visualizing Indicator

AR Visual Indicator Anchoring. Once a referent is identified through graph traversal reasoning, Speech-to-Spatial generates an AR visual indicator to situate the instruction in the physical environment. We anchor a directional arrow pointer directly above the identified object-node, ensuring that the referent is immediately visible to the remote user. This anchoring is persistent until an action (e.g., “moving”) has been performed to the referent.

Instruction Summarization and Step Ordering. On top of the visual indicator, Speech-to-Spatial overlays an AR instruction panel generated by an LLM. The utterance is summarized into a concise action description, such as “tighten the bolt” or “open the left panel”. When multiple steps are parsed from the instruction sequence, the system presents them in an alphabetically ordered list (e.g., A,B,C), guiding the user through the required operations in order. This summarization clarifies the required action with intent, reducing the cognitive burden of parsing long-utterance tasks.

Misguidance Avoidance. In cases where Speech-to-Spatial fails to interpret the instructions—involving conflicting conditions or loss of nuances in summarization, Speech-to-Spatial alternatively displays the raw transcription of the spoken instruction without the auto-generated visual pointer, as a safe fallback mechanism. This enables Speech-to-Spatial to avoid presenting any misleading visual pointers, passing control over to the users for their interpretation (maintaining user-agency). This approach preserves fidelity to the original input while ensuring that users retain control when automation falls short. This design maintains the system fail-safe.

Anchoring a visual indicator completes our “Speech-to-Graph-to-Overlay” pipeline. The arrow indicator provides spatial ground-

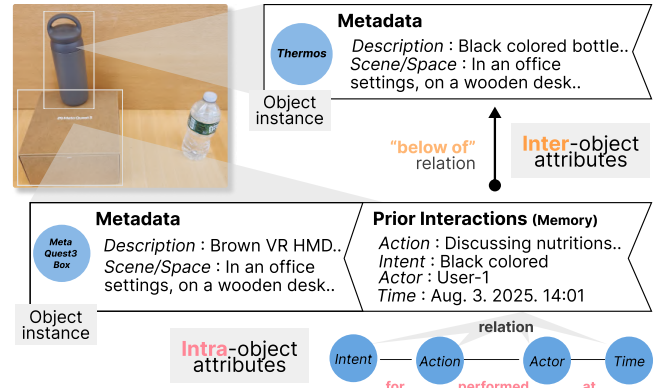


Figure 4: Object-centric relational graph: Each object maintains a graph representation that holds **Intra-object attributes**—Space, Action, Intent, Actor, Time, and Metadata—and is connected with other objects nodes—**Inter-object attributes**.

ing, while the textual overlay delivers actionable guidance. Together, they transform remote verbal instructions into a spatial indicator that disambiguates the object-of-interest and clarifies actions.

4 USE CASES AND APPLICATIONS

We illustrate how Speech-to-Spatial operates across the applications of Remote maintenance, Navigation, and Personal assistance, emphasizing its ability to provide clarity in conversations, preserve context, and reduce ambiguity in the topic of discussion. Each case demonstrates how referent disambiguation can transform under-specified speech queries into spatially guided instructions (Fig. 5).

4.1 Case Study 1: Speech-based Visual Annotation

In modern remote collaborative maintenance platforms (e.g., TeamViewer Assist AR [67], or Microsoft Dynamics 365 Remote Assist/Teams [51]), a remote expert supports the technician in the field primarily through spoken instructions while observing a shared video feed. Other platforms provide annotation tools that allow experts to label regions in the view. Yet, verbal description is the primary source of assistance, and visual annotation requires a human expert to manually indicate the target as well. With Speech-to-Spatial, an utterance (e.g., “locate the second fuse from the left, just below the green wire”) is parsed into an object-centric graph, and the remote instructor’s referent is automatically disambiguated and indicated with a visual overlay. This transformation provides the technician with unambiguous, visual directions, without requiring manual effort to visually indicate the referred target.

4.2 Case Study 2: Mapping the Speech to Visual Map

When describing the route to a destination, directions are conveyed step by step, combining landmarks and relational anchors (“walk straight until you see the entrance, and turn left”). Such instructions require the listener to maintain a mental drawing of the route, assembling each fragment into an imagined plan of the environment. While natural in daily conversations, this approach can be difficult to follow in unfamiliar or complex spaces, where the accuracy of remembering each segment is critical. Speech-to-Spatial can convert the verbal directions into a situated visual guidance using AR. Instead of relying on the user’s “mental map,” from the verbal explanation, users can visualize the directions. The wayfinding experience becomes easier to follow and transparent. The overlay makes clear what the speaker meant by “entrance” or “left,” reducing the likelihood of misunderstanding, in spoken instructions.

4.3 Case Study 3: Disambiguating the Query-of-interest

Speech-to-Spatial is useful in verbally conversing with a personal AI assistant (e.g., Gemini) as well. Here, the spoken communica-



Figure 5: Three use case scenarios of Speech-to-Spatial. (A) Remote Maintenance; (B) Indoor Navigation; and (C) Personal AI assistant.

tion can introduce ambiguity in identifying the intended referent. The user may refer to Referent A, but the AI may misinterpret it as Referent B, providing an incorrect response to the user’s query. Speech-to-Spatial can mitigate this by providing a visual indicator (e.g., a dot indicating the object-of-interest [40]) after receiving the speech input. Furthermore, this capability suggests a potential pathway towards the transformation of a customer-service AI chatbot into a visually grounded agent that can visually guide (on top of speech/textual) even with a single shared snapshot capture.

5 EVALUATION

We evaluate Speech-to-Spatial as a *referent disambiguation mechanism* that converts verbal remote instructions into spatially grounded AR guidance, by resolving the intended referent and providing a visual indicator. Our evaluation consists of two parts. (1) **End-to-end impact**: we first test *whether grounding speech-only instructions into AR guidance with Speech-to-Spatial, improves user performance compared to a conventional voice-only baseline* (Sec. 5.1) and (2) the **feasibility** of our mechanism: we analyze the spoken instructions and interaction traces logged during our open-ended study to quantify *how reliable our mechanism resolves the intended referent* (Sec. 5.2).

5.1 User Evaluation: Quantifying the Impact

5.1.1 Study Setup and Procedure

Participants. We recruited 18 participants (12 male, 6 female, aged 22-34). All participants were either Full Professional Proficiency (N=4) or Native (N=14) in English. The 7-point Likert scale indicating prior familiarity with XR systems (1:None; 7:Experienced) varied widely ($\mu=2.7$, $\sigma=1.2$), but the participants were not exposed to the task or conditions prior to the experiment. The participants (P1-P18) were compensated with a \$15 gift card.

Apparatus. Participants were given a Meta Quest3 MR HMD (passthrough mode) with its controller, to participate in the study. The HMD was to provide a controlled testbed across participants that ruled out any arm fatigue concerns (e.g., Gorilla Arm Effect) that may arise from a hand-held AR device, and allow users to perform tasks with hands (hands-free), under a practical remote assistance workflow (“receive-instructions-and-execute”). The spatial layout of the physical scene and the eight virtual cubes (mock physical referents) were registered and anchored prior to the study. To maintain uniform study conditions, the coordinate system was synchronized across sessions, and the study was conducted on the same desk settings in a lab (refer to Suppl. Material).

Procedure and Measures. Participants were given a cube interaction task under three conditions with different modalities (each modality is labeled, a ‘block’): (1) *Audio* (spoken instructions only; denoted “**Audio**”); (2) *Audio+Visual Indicator+Full transcription* (verbatim instruction displayed; denoted “**Full**”); (3) *Audio+Visual Indicator+Summarized transcription* (condensed directive above a target referent; denoted “**Summary**”). Note that the speech-only condition (1–*Audio*) is our **baseline**, and 2–*Full* and 3–*Summary*, use Speech-to-Spatial to disambiguate instructions and provide

spatial guidance (generated before the study). Within each block, participants complete four sub-blocks consisting of three trials, each covering one of four spatial referencing patterns (Sec. 3.1.1).

Every user action was logged to derive task completion time, accuracy, and interaction traces. After each sub-block, participants filled out a 0-100 scaled RTLX (Raw TLX; Unweighted NASA-TLX) and a Single Ease Question (SEQ). After each modality block, they completed another RTLX for overall experience assessment. A post-study questionnaire was given to collect overall user experience, followed by a semi-structured interview. The study conditions and tasks were presented in a counterbalanced order to mitigate sequence effects. Also, to maintain consistent motor factors (pinch, hand gestures), participants were instructed to start the task at the same designated location, and the target cubes were repositioned to maintain an equal distance of 35cm ($Distance_{hand-to-target}$). All instructions were delivered as AI-synthesized speech to avoid between-speaker prosody effects, and the audio was played at the beginning of every trial. When the sub-block referencing pattern type is “Memory-based,” users were shown to remember that this is “*the cube referred by the memory*”. Each of the eight cubes was textured uniquely. We collected a total of 1,296 trials across all participants (3 blocks \times 3 trials \times 4 sub-blocks \times 2 task types \times 18 participants), excluding a 20-minute functionality familiarization phase (tutorial).

Tasks. The tasks simulated remote guidance scenarios, where participants were verbally guided to select a target cube among distractors, with instructions balanced across reference styles (each trial is given a single task). Each participant was given equal aggregated counts of trials and tasks:

- **Locate:** Identify and select a target cube among distractors by hitting it with a virtual hammer mapped to the Quest controller. The task hypothesizes a “find” instruction, given a description (e.g., “*Locate the purple striped cube*”).
- **Move:** Move a specified cube from its current location to a designated target position on the desk. The action is performed using a pinching gesture (bare-handed). The scenario hypothesizes a remote-assisted task that involves transformation of an on-site object. The task is considered complete once a cube was selected and moved to any position (e.g., “*Move the red cube to the left of the blue dotted cube*”).

In *Audio*, instructions were delivered verbally with no additional cues mimicking traditional speech-only remote guidance. In *Full*, the spoken instruction was displayed verbatim above the target with an arrow. For *Summary*, the instruction was condensed into a concise directive and displayed alongside the arrow. We conduct a within-subject study with these hypotheses: (H1) Speech-to-Spatial will disambiguate instructions and improve clarity than the baseline; (H2) Summarized instructions will be the top choice; (H3) The advanced task (Move) will have higher demand for disambiguation.

5.1.2 Results

We report our findings at three levels computed for each Task: Overall effects across all trials, Modality/block-level comparisons, and Effects on the performance for each referencing pattern. For

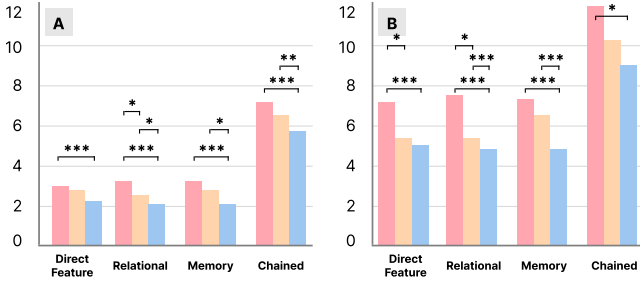


Figure 6: Comparison of median task completion time per referencing pattern : (A) Locate Task, (B) Move Task result; *** $p < .001$, ** $p < .01$, * $p < .05$; Audio, Full, Summary.

each block, we perform a repeated measures ANOVA test (when normality held), or Friedman followed by Bonferroni-corrected pairwise tests (or Wilcoxon signed-rank) with effect sizes. We use Shapiro-Wilk for normality checks. For each section and task, we summarize the insights, and provide statistical grounds behind them (as the tasks involve different levels of complexity and motor).

Task Completion Time and Accuracy. In both tasks, the disambiguated-instruction conditions (*Full*, *Summary*) reduced completion time without harming accuracy compared to *Audio*. The accuracy of Locate was near maximum under all conditions with only the visual guidance time varying. For Move, they significantly improved both speed and accuracy, with *Summary* providing the strongest benefits, and *Full* not showing a strong trend with *Audio*.

Locate: the median completion time (ascending sorted) for the conditions was: *Summary* $3.25(\pm 0.53)s$, *Full* $4.08(\pm 0.40)s$, and *Audio* $4.33(\pm 0.41)s$. Friedman test showed significance across conditions on Time ($\chi^2 = 21.00$, $p < .001$), and pairwise comparison indicated: *Summary* < *Audio* ($p < .001$), *Summary* < *Full* ($p < .001$). *Full* vs *Audio* did not show any significance. Accuracy was: *Summary* $0.998(\pm 0.001)$, *Full* $0.991(\pm 0.027)$, and *Audio* $0.986(\pm 0.032)$. No significance was found among conditions.

Move: this task requires more complex reasoning (identify and move). The mean completion time were: *Summary* $6.33(\pm 1.07)s$, *Full* $7.94(\pm 1.22)s$, and *Audio* $9.31(\pm 2.28)s$. Conditions significantly affected time ($\chi^2 = 20.33$, $p < .001$), with *Summary* < *Full* < *Audio* ($p < .05$). Accuracy of *Full* and *Summary* also improved over *Audio*: *Summary* $0.731(\pm 0.061)$, *Full* $0.676(\pm 0.137)$, and *Audio* $0.644(\pm 0.130)$, ($\chi^2 = 8.04$, $p < .018$). A Wilcoxon signed-rank test showed: *Summary* > *Audio* ($p < .020$), while no other conditions were found to be significant. This indicates that *Summary* improves both speed and accuracy.

Perceived Difficulty. Disambiguated guidance (ours) reduced perceived difficulty (higher: more difficult) robustly for the Move task, while the Locate task showed non-significance. Difficulty was measured with a 0-to-100 scale (0:Low; 100:High).

Locate: the difficulty of each condition in ascending order: *Summary* $26.39(\pm 23.01)$, *Full* $28.19(\pm 24.04)$, and *Audio* $30.97(\pm 19.61)$, did not reap any significance.

Move: showed significance (RM-ANOVA $F(2, 34) = 7.85$, $p < .05$) with *Full* vs *Audio* (post-hoc $p = .020$), and *Summary* vs *Audio* (post-hoc $p = .016$). The difficulty of each condition was: *Summary* $33.47(\pm 24.54)$, *Full* $34.72(\pm 23.13)$, and *Audio* $45.14(\pm 17.60)$.

Perceived Confidence. The perceived confidence in answers increased with *Summary* significantly higher than *Audio*, in Move. Locate did not exhibit any significant trend. Confidence was measured with a 0-to-100 scale (0:Low; 100:High).

Locate: No significance among conditions was found. Each condition scored: *Full* $89.17(\pm 13.09)$, *Summary* $88.89(\pm 11.48)$, and *Audio* $81.94(\pm 18.70)$.

Move: Confidence in the answer of each condition was:

Summary $88.89(\pm 11.12)$, *Full* $85.56(\pm 15.30)$, and *Audio* $77.64(\pm 20.89)$. Showing significance (RM-ANOVA $F(2, 34) = 5.61$, $p < .01$) for *Summary* vs *Audio* (post-hoc $p = .036$).

Cognitive Workload. Interestingly, *Full* had lower average load than *Summary* in Mental Load and Effort, across both tasks. While it did not reap any significance between *Summary* and *Full*, *Full* showed lower mean than even *Summary*. Only for the Move task did both the spatial guidance (*Full*, *Summary*) reduced mental demand as well as effort, relative to *Audio*. The Locate task did not show any statistically meaningful pattern. Each load was measured with a 0-to-100 scale (0:Low; 100:High).

Locate: Mental demand and Effort were: *Full* $23.19(\pm 23.62)$, *Summary* $27.36(\pm 25.60)$, *Audio* $30.69(\pm 20.25)$, and *Full* $22.08(\pm 19.50)$, *Summary* $25.97(\pm 22.48)$, *Audio* $29.31(\pm 18.49)$.

Move: Mental demand and Effort were: *Full* $32.22(\pm 24.16)$, *Summary* $34.31(\pm 27.93)$, *Audio* $47.22(\pm 20.45)$, and *Full* $33.06(\pm 22.70)$, *Summary* $34.31(\pm 23.79)$, *Audio* $46.53(\pm 17.62)$, respectively. It shows significance (RM-ANOVA $F(2, 34) = 7.80$, $p < .01$) on *Full* vs *Audio* (post-hoc $p = .006$) for Mental demand, and Effort also shows significance (RM-ANOVA $F(2, 34) = 9.72$, $p < .001$) in *Full* vs *Audio* (post-hoc $p = .005$), and *Summary* vs *Audio* (post-hoc $p = .013$).

Effects Per Referencing Pattern. We examine how referencing patterns (Direct Feature, Relational, Memory, Chained) shape performance within each task (Locate, Move) and by modality (*Audio*, *Full*, *Summary*). *Summary* consistently shortened completion time relative to *Audio* across all patterns and tasks, and it also showed better results than *Full* for Relational, Memory, and Chained speech patterns. The effect is largest whenever the utterance requires cross-object reasoning or recall (Memory, Chained), where the concise directive appears to reduce parsing effort while the arrow removes residual spatial ambiguity (Fig. 6). For Accuracy, Locate does not show a noticeable pattern, while for the Move task, *Summary* yields the most gains for Memory-based references. Please refer to Supplementary Material for more analysis.

Usability and Preference. Participants indicated the value of disambiguated guidance positively, rating (7-point scale) the use of *Summary*, at $5.37(\pm 1.26)$ with 79% rated higher than or equal to 5, and 53% rated higher or equal to 6. The visual reliance compared to *Audio* was 6.84 vs. 3.16 (ratio). The participants indicated that the visual anchoring reduces memory burden and expedites action “Seeing the arrows made it easier to remember the steps to be taken” (P3). When asked to rank the preferred mode of assistance, *Summary* was ranked highest (N=11), followed by *Full* (N=6), and *Audio* (N=1). “Summary captures key points without overwhelming detail, and the arrow makes instructions easy to follow” (P5). The *Audio*, which is the most common way of remote assistance in the field, was viewed as insufficient by the participants. “Audio was less helpful I need to think and remember details” (P11); “After getting used to the visual one, audio was just distracting.” (P8).

5.2 System Evaluation: Gauging the Feasibility

With the same participants post-study, we assess our framework as a whole with the participants’ own verbal input. In this session, the participants were not constrained to any specific tasks, but acted as both the remote expert (who provided the verbal instructions) and the local user (who shared the view). An example query can be: “Locate the phone to the left of the laptop” (while having a phone and a laptop in sight). This session was conducted to gauge the raw technical capabilities of our framework. The participants were instructed to provide any verbal queries and provide open-ended feedback on their user experience of Speech-to-Spatial.

We collected 81 user query instances along with qualitative feedback from the participants. Our analysis indicates that Speech-to-Spatial successfully processed 77.8% (counts: 63) of total queries,

while 13.6% (11) triggered our fall-back mechanism (*Full*). The verbal input contained incorrectly parsed noise or filler utterances (e.g., “um..”) leading to 8.6% (7) speech recognition errors. Overall, the system was commented on positively on its ability to disambiguate instructions through *Full* or *Summary* visualizations (P1,P3,P4,P9,P10). However, one participant expressed frustration with the persistent visualization of *Full* transcriptions (“*I don’t know if it’s the system or me*”) (P1). The fail-safe mechanism where the raw transcription is provided to the user upon conflicts or failures caused confusion between the system’s reliability and voice recognition errors. Another participant perceived the same fall-back behavior as beneficial, noting “*fallback helped, it shouldn’t just fail, at least it shows where it possibly went wrong.*” (P9). Speech-to-Spatial was unable to correctly parse queries involving unsupported spatial expressions. For example, multi-step spatial queries (multiple Chaining spatial references): “*second to the right of*”, “*in between object A and B*” (P2), and “*what’s next to the one on the right*” (P7). Also, a user view-oriented explanation (“*one to the right of what I’m viewing*”) (P17). We extend the discussion on these limitations in Sec. 6.

6 LIMITATIONS AND DISCUSSION

We demonstrated that our speech disambiguation framework reduces task comprehension difficulty, increases task performance, and performs with reasonable robustness. However, our system makes a few assumptions in language, sensing, and evaluation scope. Below we outline key limitations and future work. We expand our implementation details in the Supplementary Material.

Extended Evaluation Scope. We showed how graph-based referent disambiguation can convert spoken instructions into grounded guidance and improve instruction comprehension. However, to better reflect practical remote assistance scenarios, we plan to move beyond lab-controlled, single-turn prompts and synthetically generated utterances (simple tasks), which may not fully capture the variability and pragmatics of natural communication (“*Some instructions were not how I’d describe things*” (P17)—phrases becoming overly convoluted). We plan to derive instruction patterns from realistic bi-directional assistance session transcripts, and in 3D remote scenarios to observe any behavioral shifts from users, and conduct a more systematic evaluation with more participants.

Communication Pattern Coverage. We conducted the formative study to extend the established findings on spatial referencing patterns to a remote assistive scenario, and categorized the recurring patterns. However, we limit the scope of these patterns to object-centered referencing. In our studies, we found that users use view-centered (“object to my left”), or environment-centered (“near the wall”) referencing strategies as well. Furthermore, our formative study accounts for only cumulative speech patterns, instead of separating initial description patterns and recovery attempts (we only qualitatively report the orderings). We plan to expand the support of the graph representation beyond object-level, with deeper speech pattern analysis and their transition strategies. Also, we will analyze the speech patterns of users, of varying technical background.

Visual Guidance Design Space. Our core novelty lies in disambiguating referents from speech. The visual guidance is a byproduct of referent resolution. However, the effectiveness of this guidance can depend on how it is represented (dimensionality: 2D, 3D; type: arrow, circle) [25, 70]. Also, limited resolution density and graphics fidelity can reduce text legibility in AR. As a next step, we will systematically separate and evaluate the design factors, and explore richer visual encoding (e.g., icon, figures) [45], for a more informed AR guidance design and improved pipeline.

Importance of Multi-modality. Practical remote assistance situations include hands-occupied tasks (e.g., holding tools, fixing parts) where additional manual interaction is undesirable, and speech-

only becomes the most practical channel. This highlights why a speech-only mode is not simply a weaker interface but a necessary operational mode in certain cases. However, providing additional cues (e.g., gaze, gesture) strengthens the system’s user context comprehension [27, 40, 41], enabling it to better assist the user. In our future work, we will explore adaptive modality switching strategies [37] based on inferred hand availability and situational constraints.

Handling Complex Spatial Referencing. Our system showed limitations when confronted with advanced chained expressions that required multi-step reasoning such as nested or ordinal references (“*in between*”, or “*second to the left of*”). In addition, our spatial relations are limited to session axis-aligned 6DoF pose. We plan to address this with multi-hop reasoning capabilities by iteratively traversing through nodes for global relational understanding.

Robustness in the Wild. The present implementation assumes synchronized AR session coordinates and reliable sensing, conditions that are difficult to guarantee in real-world deployments. Furthermore, the spatial references can vary by the captured viewpoint. The spatial relation of object A (which is on the left of object B) can shift with the viewer’s perspective. If object A is viewed from the other side, object A is to the right of object B. These constraints challenge the scaling of our system beyond laboratory conditions. We plan to address this by introducing adaptive referencing strategies, capturing the viewpoint of the user upon relational graph initialization and applying LLM-based spatial reasoning over geometric position-based spatial layout computation.

User Agency and Transparency. Summarized guidance effectively reduced cognitive demand by distilling lengthy utterances into concise directives anchored to referents. However, condensation risks omitting semantic nuance that may be essential in complex verbal instructions (33% of the participants preferred *Full* over *Summary*). For future work, we plan to adapt Speech-to-Spatial to provide a heightened level-of-control on the visual feed of the transcription to the users, as well as providing heightened transparency into the reasoning process (e.g., failure cause). This will allow users to adjust the balance between clarity and comfort according to their task demands, and adjust verbal communication strategies.

7 CONCLUSION

We presented Speech-to-Spatial, a referent disambiguation system that converts verbal instructions into spatially grounded AR guidance for remote assistance. Motivated by our formative study showing the potential for extending recurring spatial referencing patterns to remote guidance scenarios, Speech-to-Spatial interprets how people specify targets through speech, and resolves under-specified references without relying on manual expert annotation or other cues such as gaze or gesture. We parse instructions into representative reference types and ground them in an object-centric relational graph to disambiguate the intended referent, then situate a visual indicator with a concise directive. Our evaluation showed that our pipeline makes verbal instructions easier to follow—reduced completion time, improved accuracy, lowered perceived difficulty, and workload compared to verbal-only guidance. We also demonstrated system robustness and broader applicability through case studies. We suggested that our system can be a lightweight bridge from verbal to visually explainable, actionable guidance.

8 DATA PRIVACY AND ETHICS

Participants provided informed consent prior to the study, and their identity was anonymized. Both studies were conducted under the IRB compliance of Stony Brook University (1173920).

ACKNOWLEDGMENTS

This research was supported in part by NSF award IIS2529207 and ONR award N000142312124.

REFERENCES

- [1] H. Bai, P. Sasikumar, J. Yang, and M. Billinghurst. A user study on mixed reality remote collaboration with eye gaze and hand gesture sharing. In *Proc. of ACM CHI*, pp. 1–13, 2020. 1
- [2] R. A. Bolt. “put-that-there” voice and gesture at the graphics interface. In *Proc. of SIGGRAPH*, pp. 262–270, 1980. 1, 2, 4
- [3] R. Bovo, D. Giunchi, P. Cascarano, E. J. Gonzalez, and M. Gonzalez-Franco. Revisiting put-that-there, context aware window interactions via llms. *arXiv preprint arXiv:2511.02378*, 2025. 2
- [4] M. Brehmer. Video-conferencing beyond screen-sharing and thumbnail webcam videos: Gesture-aware augmented reality video for data-rich remote presentations. *arXiv preprint arXiv:2501.05345*, 2025. 2
- [5] S. E. Brennan and H. H. Clark. Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition*, 22(6):1482, 1996. 2, 3
- [6] N. Bressa, J. Vermeulen, and W. Willett. Data every day: Designing and living with personal situated visualizations. In *Proc. of ACM CHI*, pp. 1–18, 2022. 3
- [7] N. Carbonell and S. Kieffer. Do oral messages help visual search. *Advances in natural multimodal dialogue systems*, 30:131–157, 2005. 2
- [8] R. S. M. Chan, A. Marx, A. Kim, and M. El-Assady. A design space for intelligent dialogue augmentation. In *Proc. of IUI*, pp. 18–36, 2025. 2
- [9] S. Chen, P.-L. Guhur, M. Tapaswi, C. Schmid, and I. Laptev. Language conditioned spatial relation reasoning for 3d object grounding. *Neurips*, 35:20522–20535, 2022. 2
- [10] H. H. Clark and S. E. Brennan. Grounding in communication. In L. Resnick, L. B., M. John, S. Teasley, and D., eds., *Perspectives on Socially Shared Cognition*, pp. 13–1991. APA, 1991. 3
- [11] F. I. Doğan, S. Kalkan, and I. Leite. Learning to generate unambiguous spatial referring expressions for real-world environments. In *Proc. of IEEE/RSJ IROS*, pp. 4992–4999, 2019. 2
- [12] M. D. Dogan, E. J. Gonzalez, K. Ahuja, R. Du, A. Colaço, J. Lee, M. Gonzalez-Franco, and D. Kim. Augmented object intelligence with XR-Objects. In *Proc. of ACM UIST*, pp. 1–15, 2024. 2
- [13] D. dos Santos Silva and I. Paraboni. Generating spatial referring expressions in interactive 3d worlds. *Spatial Cognition & Computation*, 15(3):186–225, 2015. 2
- [14] R. Druta, C. Druta, P. Negirla, and I. Silea. A review on methods and systems for remote collaboration. *Applied Sciences*, 11(21):10035, 2021. 1, 2
- [15] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, and J. Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv:2404.16130*, 2024. 2, 5
- [16] A. Evgrashin. Whisper for unity. <https://github.com/Macoron/whisper.unity/tree/master>, 2024. Aug. 31. 2024. 4
- [17] C. G. Fidalgo, Y. Yan, H. Cho, M. Sousa, D. Lindlbauer, and J. Jorge. A survey on remote assistance and training in mixed reality environments. *IEEE TVCG*, 29(5):2291–2303, 2023. 2
- [18] D. I. Fink, J. Zagermann, H. Reiterer, and H.-C. Jetter. Re-locations: Augmenting personal and shared workspaces to support remote collaboration in incongruent spaces. *Proc. of ACM HCI*, 6(ISS):1–30, 2022. 2
- [19] A. Garnham. A unified theory of the meaning of some spatial relational terms. *Cognition*, 31(1):45–60, 1989. 3
- [20] J. E. S. Grønbaek, K. Pfeuffer, E. Velloso, M. Astrup, M. I. S. Pedersen, M. Kjær, G. Leiva, and H. Gellersen. Partially blended realities: Aligning dissimilar spaces for distributed mixed reality meetings. In *Proc. of ACM CHI*, pp. 1–16, 2023. 2
- [21] J. Grubert, T. Langlotz, S. Zollmann, and H. Regenbrecht. Towards pervasive augmented reality: Context-awareness in augmented reality. *IEEE TVCG*, 23(6):1706–1724, 2016. 3
- [22] Q. Gu, A. Kuwajerwala, S. Morin, K. M. Jatavallabhula, B. Sen, A. Agarwal, C. Rivera, W. Paul, K. Ellis, R. Chellappa, et al. Concept-graphs: Open-vocabulary 3d scene graphs for perception and planning. In *Proc. of IEEE ICRA*, pp. 5021–5028, 2024. 2, 5
- [23] P. Gurevich, J. Lanir, B. Cohen, and R. Stone. Teleadvisor: a versatile augmented reality tool for remote assistance. In *Proc. of ACM CHI*, pp. 619–622, 2012. 2
- [24] C. Han and K. E. Isaacs. A deixis-centered approach for documenting remote synchronous communication around data visualizations. *IEEE TVCG*, 2024. 2
- [25] D. Hepperle, Y. Weiß, A. Siess, and M. Wölfel. 2d, 3d or speech? a case study on which user interface is preferable for what kind of object interaction in immersive virtual reality. *Computers & Graphics*, 82:321–331, 2019. 9
- [26] P. Howlader, H. Nguyen-Canh, S. Das, J. Xu, H. Le, and D. Samaras. Cora: Consistency-guided semi-supervised framework for reasoning segmentation. In *Proc. of IEEE/CVF WACV*, 2026. 2
- [27] X. Hu, D. Ma, F. He, Z. Zhu, S.-K. Hsia, C. Zhu, Z. Liu, and K. Ramani. Gsprompt: Leveraging co-speech gestures to augment llm-based interaction in virtual reality. In *Proc. of ACM DIS*, pp. 59–80, 2025. 1, 2, 4, 9
- [28] S. Jadon, M. Faridan, E. Mah, R. Vaish, W. Willett, and R. Suzuki. Augmented conversation with embedded speech-driven on-the-fly referencing in ar. *arXiv preprint arXiv:2405.18537*, 2024. 1, 2
- [29] S. Jang, E.-J. Ko, and W. Woo. Unified user-centric context: Who, where, when, what, how and why. In *Proc. of UbiPCMM*, 2005. 4, 5
- [30] K. Johannsen and J. P. D. Ruiters. Reference frame selection in dialog: priming or preference? *Frontiers in Human Neuroscience*, 7:667, 2013. 2, 3
- [31] R. Kartmann and T. Asfour. Interactive and incremental learning of spatial object relations from human demonstrations. *Frontiers in Robotics and AI*, 10:1151303, 2023. 2
- [32] D. Kim, T. Ha, J. Hong, S. Kim, S. Choi, H. Ko, and W. Woo. Meta-objects: Interactive and multisensory virtual objects learned from the real world for use in augmented reality. *IEEE CG&A*, 45(3):134–143, 2025. 2
- [33] H. Kim, E. Hu, and S. Heo. Spaceshare: Leveraging multimodal context for fluid sharing of spaces in video meetings. In *Proc. of ACM UIST-Adjunct*, pp. 1–3, 2025. 2
- [34] H. Kim, T. Matuszka, J.-I. Kim, J. Kim, and W. Woo. Ontology-based mobile augmented reality in cultural heritage sites: information modeling and user study. *Multimedia Tools and Applications*, 76(24):26001–26029, 2017. 1, 5
- [35] Y. Kim, Z. Aamir, M. Singh, S. Boorboor, K. Mueller, and A. E. Kaufman. Explainable xr: Understanding user behaviors of xr environments using llm-assisted analytics framework. *IEEE TVCG*, 2025. 4, 5
- [36] B. Lee, M. Sedlmair, and D. Schmalstieg. Design patterns for situated visualization in augmented reality. *IEEE TVCG*, 30(1):1324–1335, 2023. 3
- [37] G. Lee, M. Xia, N. Numan, X. Qian, D. Li, Y. Chen, A. Kulshrestha, I. Chatterjee, Y. Zhang, D. Manocha, et al. Sensible agent: A framework for unobtrusive interaction with proactive ar agents. In *Proc. of ACM UIST*, pp. 1–22, 2025. 9
- [38] J. Lee, F. Aleotti, D. Mazala, G. Garcia-Hernando, S. Vicente, O. J. Johnston, I. Kraus-Liang, J. Powierza, D. Shin, J. E. Froehlich, et al. Imaginear: Ai-assisted in-situ authoring in augmented reality. In *Proc. of ACM UIST*, pp. 1–21, 2025. 2
- [39] J. Lee, J. Kim, J. Ahn, and W. Woo. Remote diagnosis of architectural heritage based on 5w1h model-based metadata in virtual reality. *ISPRS IJGI*, 8(8):339, 2019. 4
- [40] J. Lee, J. Wang, E. Brown, L. Chu, S. S. Rodriguez, and J. E. Froehlich. GazePointAR: a context-aware multimodal voice assistant for pronoun disambiguation in wearable augmented reality. In *Proc. of ACM CHI*, pp. 1–20, 2024. 1, 2, 4, 7, 9
- [41] J. Lee, T. Wang, J. Fashimpaur, N. Sendhilnathan, and T. R. Jonker. Walkie-talkie: Exploring longitudinal natural gaze, llms, and vlms for query disambiguation in xr. In *Proc. of ACM CHI EA*, pp. 1–9, 2025. 1, 2, 9
- [42] W. J. Levelt. Cognitive styles in the use of spatial direction terms. *Psychology*, 1982. 2, 3
- [43] W. J. Levelt. *Speaking: From intention to articulation*. MIT press, 1993. 2, 3
- [44] S. C. Levinson. Frames of reference and molyneux’s question: Crosslinguistic evidence. *Language and space*, 109:169, 1996. 3
- [45] C. Li, G. Wu, G. Y.-Y. Chan, D. G. Turakhia, S. Castelo Quispe, D. Li,

- L. Welch, C. Silva, and J. Qian. Satori: Towards proactive ar assistant with belief-desire-intention user modeling. In *Proc. of ACM CHI*, pp. 1–24, 2025. 2, 9
- [46] J. N. Li, Z. Zhang, and J. Ma. Omniquery: Contextually augmenting captured multimodal memories to enable personal question answering. In *Proc. of ACM CHI*, pp. 1–20, 2025. 3
- [47] X. Liu, D. Jia, X. C. Liu, M. Gonzalez-Franco, and C. Zhu-Tian. Reality proxy: fluid interactions with real-world objects in mr via abstract representations. In *Proc. of ACM UIST*, pp. 1–16, 2025. 2
- [48] E. Lukianova, J.-Y. Jeong, and J.-W. Jeong. A picture is worth a thousand words? investigating the impact of image aids in ar on memory recall for everyday tasks. In *Proc. of IUI*, pp. 106–126, 2025. 3
- [49] M. N. Lystbæk, K. Pfeuffer, T. Langlotz, J. E. S. Grønbaek, and H. Gellersen. Spatial gaze markers: Supporting effective task switching in augmented reality. In *Proc. of ACM CHI*, pp. 1–11, 2024. 3
- [50] D. Markov-Vetter, M. Luboschik, A. T. Islam, P. Gauger, and O. Staadt. The effect of spatial reference on visual attention and workload during viewpoint guidance in augmented reality. In *Proc. of ACM SUI*, pp. 1–10, 2020. 2
- [51] Microsoft. Dynamics 365 remote assist. <https://learn.microsoft.com/en-us/dynamics365/mixed-reality/remote-assist/ra-overview>, 2025. Sep. 3. 2025. 6
- [52] G. A. Miller and P. N. Johnson-Laird. *Language and perception*. Harvard University Press, 1976. 3
- [53] R. Murai, E. Dexheimer, and A. J. Davison. Mast3r-slam: Real-time dense slam with 3d reconstruction priors. In *Proc. of CVPR*, pp. 16695–16705, 2025. 1, 2, 5
- [54] M. Rebol, C. Hood, C. Ranniger, A. Rutenberg, N. Sikka, E. M. Horan, C. Gütl, and K. Pietroszek. Remote assistance with mixed reality for procedural tasks. In *Proc. of IEEE VRW*, pp. 653–654, 2021. 2
- [55] A. Rosinol, A. Gupta, M. Abate, J. Shi, and L. Carlone. 3d dynamic scene graphs: Actionable spatial perception with places, objects, and humans. *arXiv preprint arXiv:2002.06289*, 2020. 2
- [56] K. A. Satriadi, B. Tag, and T. Dwyer. Context-dependent memory in situated visualization. *arXiv:2311.12288*, 2023. 3
- [57] M. F. Schober. Addressee-and object-centered frames of reference in spatial descriptions. In *American Association for Artificial Intelligence, Working Notes of the 1996 AAAI Spring Symposium on Cognitive and Computational Models of Spatial Representation*, vol. 47, pp. 92–100, 1996. 2, 3
- [58] S. Schütz, A. Gatt, and S. Zarriß. Rethinking symbolic and visual context in referring expression generation. *Frontiers in Artificial Intelligence*, 6:1067125, 2023. 2
- [59] J. Seo, I. Avellino, D. P. Rajasagi, A. Komlodi, and H. M. Mentis. Holomenter: Enabling remote instruction through augmented reality mobile views. *Proc. of ACM HCI*, 7(GROUP):1–29, 2023. 2
- [60] M. Shakeri, H. Park, I. Jeon, A. Sadeghi-Niaraki, and W. Woo. User behavior modeling for ar personalized recommendations in spatial transitions. *VR*, 27(4):3033–3050, 2023. 5
- [61] J. Shen, J. J. Dudley, and P. O. Kristensson. Encode-store-retrieve: Augmenting human memory through language-encoded egocentric perception. In *Proc. of IEEE ISMAR*, pp. 923–931, 2024. 3
- [62] A. Shusterman and P. Li. Frames of reference in spatial language acquisition. *Cognitive psychology*, 88:115–161, 2016. 2, 3
- [63] J. G. R. d. Souza, J. J. Ferreira, and V. Segura. A taxonomy of methods, tools, and approaches for enabling collaborative annotation. In *Proc. of IHC*, pp. 1–12, 2023. 2
- [64] D. Stover and D. Bowman. Taggar: General-purpose task guidance from natural language in augmented reality using vision-language models. In *Proc. of ACM SUI*, pp. 1–12, 2024. 1, 2
- [65] H. A. Taylor and B. Tversky. Descriptions and depictions of environments. *Memory & cognition*, 20(5):483–496, 1992. 3
- [66] H. A. Taylor and B. Tversky. Perspective in spatial descriptions. *Journal of memory and language*, 35(3):371–391, 1996. 2, 3
- [67] TeamViewer. Teamviewer assist ar. <https://www.teamviewer.com/en-us/products/frontline/solutions/remote-assistance>, 2025. Sep. 3. 2025. 6
- [68] P. Wang, Y. Wang, Y. Wang, M. Billinghamurst, D. Yang, H. Yang, R. Luo, and X. Zhang. Extended reality remote collaboration supporting visual annotation cues for industry: A literature review. *Engineered Science*, 37:1802, 2025. 2
- [69] F. Zaman, C. Anslow, and T. J. Rhee. Vicarious: Context-aware viewpoints selection for mixed reality collaboration. In *Proc. of ACM VRST*, pp. 1–11, 2023. 2
- [70] A. Y. Zhao, A. Gunturu, E. Y.-L. Do, and R. Suzuki. Guided reality: Generating visually-enriched ar task guidance with llms and vision models. *arXiv preprint arXiv:2508.03547*, 2025. 2, 9
- [71] Zoom. Zoom. <https://www.zoom.com/>, 2025. Sep. 3. 2025. 3
- [72] W. D. Zulfikar, S. Chan, and P. Maes. Memoro: Using large language models to realize a concise interface for real-time memory augmentation. In *Proc. of ACM CHI*, pp. 1–18, 2024. 3