Virtual Minds, Real Work: LLM-Powered Preference-Based Planning through Spatial Multi-Agent-Human Collaboration

Ziyi Zhang
Southeast University
Nanjing, China
School of Information Sciences
University of Illinois at Urbana-Champaign
Champaign, Illinois, USA
ziyiz13@illinois.edu

Xin Yi
Institute for Network Sciences and Cyberspace
Tsinghua University
Beijing, China
yixin@tsinghua.edu.cn

Zitong Dai
Southeast University
Nanjing, China
220245433@seu.edu.cn

Xuewen Yu
Southeast University
Nanjing, China
xwyu@seu.edu.cn

Shuning Zhang
Institute for Network Sciences and Cyberspace
Tsinghua University
Beijing, China
zsn23@mails.tsinghua.edu.cn

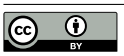
Bo Liu
Southeast University
Nanjing, China
bliu@seu.edu.cn

Jiuxin Cao
Southeast University
Nanjing, China
jx.cao@seu.edu.cn

Hantao Zhao
Southeast University
Nanjing, China
Purple Mountain Laboratories
Nanjing, China
htzhao@seu.edu.cn

Abstract

People frequently face preference-based planning tasks requiring balancing goals with nuanced constraints, yet even advanced LLMs demand considerable effort to produce and adjust plans reflecting complex user preferences. We present MAVIS (Multi-Agent Virtual Interactive Synergy), a multi-agent system within a virtual workspace that introduces an incremental collaboration mechanism. This mechanism automatically decomposes tasks into guidelines and sequentially introduces expert agents. Each agent proactively engages users in focused dialog to uncover implicit preferences, while successive agents add perspectives and transparently negotiate trade-offs. To mitigate textual overload, MAVIS employs spatial visualizations that externalize agents' reasoning through step-linked summaries and context-aware boards, with embodied avatars supporting natural interaction. Across studies, Study 1 showed our collaboration mechanism doubled expressed preferences and improved planning quality by 60.3% over a conventional LLM baseline. Study 2 affirmed visualization's benefits over a non-spatial baseline, while Study 3 confirmed its versatility across VR and desktop modalities and diverse tasks.



CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791926

CCS Concepts

• **Human-centered computing** → **Interactive systems and tools**.

Keywords

Preference-based Planning, Virtual Reality, Large Language Models, Multi-agent Systems, Human-AI Interaction, Interactive Planning, Natural Language Interaction, User-centered Design

ACM Reference Format:
Ziyi Zhang, Xin Yi, Zitong Dai, Xuewen Yu, Shuning Zhang, Bo Liu, Jiuxin Cao, and Hantao Zhao. 2026. Virtual Minds, Real Work: LLM-Powered Preference-Based Planning through Spatial Multi-Agent-Human Collaboration. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3772318.3791926

1 Introduction

Preference-based planning (PBP) refers to creating plans that achieve the desired goals and fulfill user-defined preferences as comprehensively as possible [39]. This type of planning exists in various domains, ranging from everyday decision-making [69] to more complex, professional contexts [39], such as planning a personalized vacation itinerary or coordinating a project that meets diverse stakeholder needs. By integrating user preferences organically, PBP enhances the overall usability of planning results [19] and increases user satisfaction [39]. However, the process of planning can be

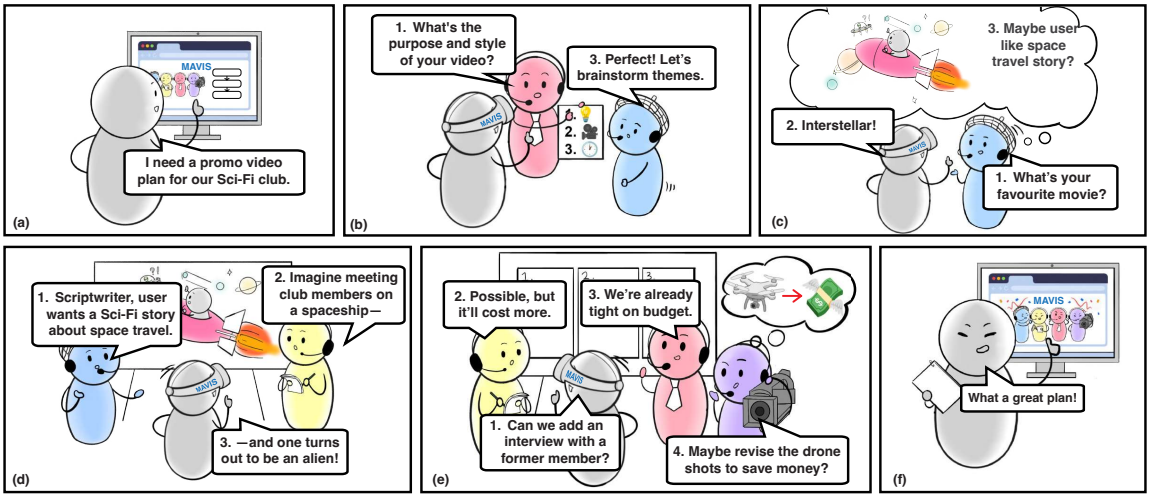


Figure 1: Workflow of MAVIS in a video production scenario. (a) MAVIS analyzes the user's input and generates a staged guideline for agent roles and entry. (b) In the virtual workspace, the Project Manager initiates collaboration and introduces Concept Manager to ground the discussion. (c) The Concept Manager iteratively elicits user preferences through conversational dialogue. (d) Additional agents join incrementally for cross-agent coordination. (e) Budget and timeline constraints are negotiated, with cascading updates applied when the user revises requirements. (f) The final preference-aligned plan is exported.

highly complex [30], often requiring careful consideration of various preferences and constraints [56]. As a result, users frequently struggle to generate optimal plans on their own.

To alleviate these difficulties, researchers have developed various planning systems [68], demonstrating their effectiveness in certain tasks [23, 33, 56]. Despite these advancements, existing planning assistance systems typically fall short in two key areas. First, most systems offer static, independent planning [1], preventing users from influencing plans during generation [68]. Since preferences typically represent soft constraints that cannot all be simultaneously satisfied, planning quality improves with user involvement [39]. Current systems lack the iterative, adaptive interactions necessary for users to dynamically adjust preferences, ultimately limiting the creation of high-quality, personalized plans. Second, existing systems often require well-structured input to define the planning task explicitly [15]. This includes detailed specifications such as the possible states of the world, available actions, initial conditions, and desired goals [20]. The requirement to represent information in specific formats (e.g., Planning Domain Definition Language-PDDL) [32] limits these systems' applicability to real-world tasks, which often involve qualitative decisions that resist precise definition [19].

Recent advances in Large Language Models (LLMs) offer potential for natural language interaction within these systems [17, 41, 58], yet they still rely on clearly articulated prompts [48, 87], struggle to capture multidimensional and evolving user preferences such as trade-offs or competing priorities [71, 85]. Emerging multi-agent LLM approaches have begun to address task complexity, yet they typically demand predefined agent roles and processes, operate through closed internal dialogues [52], offer limited transparency or opportunities for user intervention, and often produce static outputs with limited opportunities for users to refine or steer plans once execution begins [34, 51, 61]. There remains limited understanding of the practical challenges users encounter in preference-based planning and the design principles needed to effectively address them. While planning algorithms and LLMs have demonstrated technical potential, little is known about how users experience these systems in practice or which obstacles most constrain their effectiveness. Without such insights, it is difficult to move beyond static assistance toward genuinely interactive preference-based planning.

To bridge this gap, our work is guided by the following research questions:

- **RQ1:** What practical challenges do users encounter when performing complex preference-based planning with existing tools and practices, and what design principles do these challenges suggest for interactive, LLM-based planning support?
- **RQ2:** Building on these principles, how can roles, workflows, and interaction patterns be structured and coordinated in a multi-agent human–AI collaboration mechanism to better support preference-based planning?
- **RQ3:** How can visualization mechanisms and their interface modalities (e.g., immersive VR vs desktop) be designed to help users manage the information complexity of preference-based planning?

To address RQ1, we conducted a requirements elicitation study with participants who had prior experience in organizing complex, preference-driven planning tasks. The study revealed recurring difficulties in structuring tasks, balancing interdependent constraints, and managing large volumes of information. These insights informed a set of design principles that guided the development of our system.

Drawing on these insights and responding to RQ2 and RQ3, we propose MAVIS (Multi-Agent Virtual Interactive Synergy), a multi-agent system for preference-based planning in virtual workspace.

MAVIS introduces an incremental multi-agent-human collaboration mechanism that differs from conventional multi-agent systems. Rather than overwhelming users with simultaneous agent interactions or static solutions, MAVIS automatically decomposes tasks into a staged guideline and then sequentially introduces expert agents. Each agent engages users in focused dialogue to elicit specific preferences in depth, while subsequent agents contribute new perspectives, negotiate trade-offs, and propagate adjustments across the plan. Users can observe these negotiations or intervene at any point, balancing system autonomy with user control.

This mechanism is realized through two components. *Taskformer* operationalizes the staged guideline by turning a brief task description into an explicit workflow that assigns roles, schedules agent entry, and governs cross-agent negotiation and cascading updates; in contrast to typical LLM multi-agent setups with hand-crafted structures or opaque internal chats, it makes collaboration inspectable and lets the user act as a meta-planner rather than a prompt engineer. *CoNavigator* is a spatial visualization that segments extensive text into step-linked summaries and persistent context panels and embodies agents as distinct avatars so that their roles and contributions remain easy to follow, and can be instantiated in both immersive VR and desktop.

We evaluate MAVIS through three complementary evaluation studies. Study 1 quantitatively compares Taskformer with a conventional LLM baseline on travel planning, showing that incremental multi-agent collaboration increases preference articulation and plan quality without additional interactional overhead (RQ2). Study 2 qualitatively compares MAVIS (with CoNavigator) to a non-spatial web-based multi-agent baseline to understand how spatial visualization and embodiment affect engagement, sensemaking, and perceived effectiveness (RQ3). Study 3 extends this analysis to domain experts across diverse real-world planning scenarios, using a within-subjects comparison of MAVIS (VR) and MAVIS-Desktop to examine cross-task versatility and how interface modality shapes usability, engagement, and perceived suitability for everyday use (RQ3). Together, these studies disentangle the contributions of MAVIS's core mechanisms while validating its effectiveness across diverse tasks and interaction modalities.

Our contributions are threefold. First, we characterize the core challenges of preference-based planning in practice and derive design principles for interactive, LLM-based planning support. Second, we present MAVIS, which combines an incremental multi-agent collaboration mechanism (Taskformer) with spatial visualizations (CoNavigator) instantiated across VR and desktop. Third, through three studies, we provide empirical evidence that these designs can improve preference articulation and plan alignment, support more manageable sensemaking, and generalize across tasks and modalities, yielding design implications for future human-centered planning systems.

2 Related Works

2.1 Computer Assisted Preference-based Planning

Preference-based planning refers to the generation of plans that achieve specified goals while satisfying user-defined preferences as comprehensively as possible [39]. This approach is prevalent across various everyday [39] and professional contexts [39], such as arranging travel itineraries [69] or coordinating logistics [4]. Prior studies have demonstrated that by integrating user preferences organically, PBP enhances the overall usability of planning results [19] and increases user satisfaction [39]. However, due to the inherent complexity of planning processes [30], particularly when balancing multiple competing preferences [56], users often struggle to independently produce optimal outcomes.

To alleviate these challenges, researchers have developed planning systems intended to facilitate the planning process [68]. These systems typically accept clearly defined inputs, including initial conditions, available actions, and desired outcomes [20], represented in structured languages such as the PDDL [32]. Upon receiving this structured input, the planning systems employ various computational approaches [22, 35, 46] to produce actionable plans.

However, existing systems face two main limitations: the complexity of input requirements and the rigidity of planning interactions [1]. First, the necessity for strictly structured and explicitly defined task representations limits their applicability [15]. Because many real-world tasks are qualitative and heavily context-dependent, articulating them precisely using rigid languages like PDDL remains challenging [19]. In addition, converting complex contextual information into structured representations requires specialized knowledge and effort beyond typical user capabilities [68]. Moreover, current planning systems provide limited support for expressing nuanced, user-centric preferences. Although extensions like PDDL3.0 [23] introduce constructs for specifying preferences, they primarily address objective temporal conditions and inadequately represent subjective, nuanced constraints [24, 69]. Second, current planning systems predominantly perform static, independent planning processes [1, 68]. Given that user preferences inherently function as soft constraints [39], which might not be simultaneously satisfiable, user involvement and iterative adjustments during the planning process become essential [39].

Aligned with the broader aim of assisting users in preference-based planning, our work seeks to address these limitations. Our approach prioritizes intuitive, accessible user interactions, enabling users, including those lacking technical expertise, to actively engage through natural language dialogues. This allows for ongoing clarification, dynamic user interventions, and proactive system-human collaboration, ultimately improving planning quality and user satisfaction.

2.2 Large Language Models as Planner

The emergence of LLMs has introduced new opportunities for computer-assisted planning [17, 41, 58, 70, 90, 91], largely due to their capacity for leveraging commonsense knowledge [16] without requiring extensive domain input [41, 91]. Researchers have shown that LLMs can solve PDDL-defined tasks or guide heuristic planners [66], and even serve as intermediaries connecting structured inputs to human or automated feedback [27]. Yet, despite these advances, most approaches still demand structured, explicitly defined inputs, limiting the articulation of subjective or qualitative preferences [40].

A distinctive strength of LLMs is their conversational ability, which enables more natural engagement with users [42]. Studies have demonstrated their potential to process subjective and context-rich inputs across domains [5, 79, 84]. However, this potential depends heavily on carefully designed prompts [64, 82, 83], and LLMs still struggle with multidimensional user preferences such as trade-offs or competing priorities [3, 55, 71, 85].

To overcome these challenges, multi-agent LLM systems have emerged to coordinate specialized agents on complex tasks [52]. Open-source frameworks such as AutoGen [81] and CrewAI[1] provide flexible orchestration of specialized agents, but typically require developers to manually define agent roles and relationships, and demand additional custom design to include the user in the loop, which raises the barrier for non-technical users and hinders the elicitation and integration of nuanced preferences during the process. Commercial systems like OpenAI's Deep Research[2] and Google's Gemini Deep Research[3] extend multi-step planning into web browsing and synthesis, yet they confine user input to the initial query and proceed autonomously, preventing ongoing preference refinement. Academic systems such as MetaGPT [34], ChatDev [61], and CAMEL [51] showcase effective division of labor through role-playing or assembly-line protocols, but these processes largely unfold as closed internal dialogues: users cannot easily observe, intervene, or redirect intermediate steps.

Taken together, these lines of work highlight the promise of LLM-based agent collaboration but reveal a gap in interactive, user-steerable preference-based planning: systems that not only decompose and optimize plans, but also keep users in the loop as active partners who can surface evolving preferences and guide negotiations in real time. Our work responds to this gap by exploring a multi-agent–human collaboration approach that supports natural expression of vague or evolving preferences and makes agent cooperation visible and steerable during the planning process.

2.3 Spatial and Embodied Visualization for Multi-Agent-Human Collaboration

While multi-agent LLMs show promise for complex planning [52], integrating humans into the loop creates interface challenges [6]. As multiple agents simultaneously generate plans, exchange information, and surface contextual details, they can quickly overwhelm users with high volumes of overlapping text [72], leading to two critical usability barriers: *information overload*, which increases cognitive effort and undermines sensemaking [74], and *role ambiguity*, where agents' responsibilities and negotiations become opaque [11, 89], making it difficult to follow parallel negotiations or decide whom to address [11, 89]. Addressing these requires interfaces that make multi-agent collaboration both manageable and transparent.

To manage information overload, one approach is leveraging embodiment, giving each agent a humanized avatar to improve distinguishability, trust, and engagement [37, 45, 54, 75]. Embodied agents can invoke social attribution, helping users map information and actions to specific actors rather than an amorphous system [37, 75]. However, prior research largely focuses on one-to-one human–agent settings [25, 49, 76, 83], leaving open how embodiment supports teams of collaborating agents, where the goal is to understand differentiated roles, responsibilities, and negotiations rather than build a relationship with a single agent.

A complementary strategy is spatial visualization to organize and offload complex information [31]. Studies show that spatial structure can reduce memory burden and support sensemaking by enabling users to rely on spatial perception rather than mentally juggling all details [2, 63]. 3D environments can extend spatial offloading [78, 88], but their benefits hinge on structure: well-organized 3D layouts can reduce cognitive load and support creativity [26, 29, 80], whereas poorly structured designs may increase load and disorientation [12].

For collaborative contexts, immersive VR may further strengthen these effects by increasing presence and social presence [53], which are associated with higher engagement and sustained attention [57, 67]. At the same time, VR represents one end of a broader modality design space [59]: it offers richer spatial and embodied cues, but can impose physical overhead [65]. This motivates studying VR as one instantiation of spatial/embodied visualization and comparing it against a traditional desktop counterpart to understand how interface modality shapes support for such planning tasks.

Building on these findings, our work explores the integration of structured spatial visualization and embodiment to support multi-agent preference-based planning tasks. We designed a mechanism to address the cognitive burdens of multi-agent planning and evaluate its user experience impact and how interface modality (immersive VR vs. traditional desktop) influences its effectiveness.

3 Requirements Elicitation Study: Understanding Challenges in Preference-Based Planning

To clarify practical user needs and identify key challenges inherent in preference-based planning, we conducted a requirements elicitation study with participants who had prior experience in organizing complex, preference-driven planning tasks. This study aimed to (1) understand users' current practices in solving preference-based planning tasks, identifying existing gaps, challenges, and unmet needs; and (2) gather insights into potential improvement strategies through a design elicitation approach, thereby informing the subsequent design process from experienced users' perspectives.

3.1 Method

3.1.1 Participants. We recruited seven experienced participants (4 males, 3 females; $M_{age} = 22.7$, $SD = 1.67$) from university student communities, each with substantial experience in preference-based planning tasks. Two participants (E1, E2) were members of a university debate team with over three years of experience preparing sophisticated arguments. Three participants (E3–E5) had recently organized group travel itineraries for more than five individuals. Two participants (E6, E7) had independently managed multiple international trips in the past two years. These tasks were selected for their reliance on user-defined preferences and inherent complexity. Each study session lasted approximately one hour, and participants were compensated 50 CNY.

[1] https://crewai.com/
[2] https://openai.com/index/introducing-deep-research/
[3] https://gemini.google/overview/deep-research/

3.1.2 Protocol. After obtaining informed consent, participants were introduced to the study's objectives. Sessions combined two activities: (1) an open-ended discussion exploring participants' practices, challenges, and unmet needs in preference-based planning, and (2) a design elicitation activity where participants selected a personally relevant yet unsolved planning task. Using GPT-4o as a probe system, participants engaged in think-aloud interaction while articulating their reasoning processes. Following the activity, they reflected on encountered limitations and suggested improvements.

All sessions were audio-recorded, transcribed, and analyzed using Braun and Clarke's thematic analysis approach [8]. We combined inductive and deductive coding, with initial codes developed by the lead researcher and iteratively refined through team discussions [14]. The analysis yielded 35 codes grouped into 24 subthemes and 5 main themes, capturing both the challenges participants experienced and the improvements they envisioned for system support in preference-based planning.

3.2 Results

The study revealed a set of recurring, interrelated challenges in preference-based planning, together with suggestions for how systems could better support users.

Theme 1: Task oversight and the need for decomposition. All participants reported difficulty anticipating all relevant aspects of a planning task at the outset. For instance, E2 struggled to foresee possible argument angles in debate preparation, and E6 overlooked timing dependencies when organizing travel. Four participants (4/7) described concrete oversights that only surfaced during execution, forcing ad-hoc adjustments. To mitigate this, five participants (5/7) called for systems that provide incremental goal-setting guidance, progressively revealing details and prompting users to refine objectives step by step. Such staged support was seen as key to reducing cognitive burden and preventing critical elements from being missed.

Theme 2: Limited expertise and the need for proactive, iterative interaction. Three participants (3/7) highlighted limited practical and interdisciplinary knowledge as barriers to effective planning. E7 initially struggled to identify critical elements in travel planning, while E2 found it hard to construct robust arguments on unfamiliar topics, leading to uncertainty about priorities and strategy. All participants criticized existing tools as reactive and response-driven, requiring detailed initial input without support. They expressed a strong preference for systems that proactively ask clarifying questions, surface implicit preferences, and support iterative refinement without regenerating entire plans, particularly for users lacking deep domain expertise.

Theme 3: Balancing progress across multiple dimensions. Managing interdependent constraints was a universal challenge (7/7). E7 frequently underestimated budgets and logistics, causing deviations during execution; E4 and E5 described difficulties in coordinating activities with accommodation and transportation in group travel planning; E2 highlighted the complexity of balancing multiple thematic angles in debate preparation. Four participants (4/7) felt that existing tools either forced them to handle all dimensions at once, causing overload, or to focus narrowly on a single aspect, risking loss of global oversight. They argued that systems should help integrate multiple perspectives stepwise, make dependencies explicit, and support monitoring and adjusting the balance among dimensions over time.

Theme 4: Trade-offs and adjustment costs. Five participants (5/7) reported that adapting plans to new constraints or changed preferences was costly. Small modifications, such as rescheduling a single activity, often triggered cascading changes that disrupted logistics and required extensive manual re-coordination (E3, E5). This high adjustment cost discouraged exploration of alternatives. Four participants (4/7) therefore requested mechanisms for low-cost adjustment: being able to test options, flexibly adjust trade-offs, and have necessary updates automatically propagated across related components rather than editing each part by hand.

Theme 5: Cognitive overload and information management. All participants noted being overwhelmed by the volume and form of information. E4 reported feeling paralyzed by both excessively detailed and overly general content. While participants valued collaborative discussion for surfacing new ideas, they found existing tools ill-suited for integrating diverse perspectives without overload. Four participants (4/7) explicitly asked for clearer, more intuitive information presentation, favoring summaries, visual displays, and conversational formats that resemble natural dialogue. They believed such designs would sustain engagement, reduce cognitive load, and make different perspectives more distinguishable and actionable.

Collectively, these five themes reveal recurring pain points in preference-based planning and point to directions for system support. Building on these empirical insights, we derived the following **Design Principles (DPs)**, which translate participants' challenges and suggestions into actionable guidelines:

- **DP1: Support task decomposition and incremental guidance.** Systems should help users break down complex tasks into manageable subtasks and provide stepwise guidance, enabling progressive plan evolution without requiring users to foresee all details upfront.
- **DP2: Enable proactive and iterative preference elicitation.** Rather than relying solely on user-initiated input, systems should proactively prompt clarifications, uncover implicit preferences, and support refinement through iterative dialogues.
- **DP3: Provide multi-perspective integration with transparent optimization.** Systems should introduce perspectives in a structured manner, making trade-offs and dependencies explicit to help users maintain global awareness while exploring local details.
- **DP4: Facilitate flexible adjustments and trade-off exploration.** Systems should support exploratory modifications that automatically propagate across interdependent components, allowing users to explore alternatives and negotiate trade-offs with low overhead.
- **DP5: Reduce cognitive burden through visualization.** Systems should move beyond text-heavy outputs and help users process information intuitively, distinguish perspectives more clearly, and remain actively engaged without being overwhelmed.

4 The MAVIS System

Informed by the requirements elicitation study and the five design principles (DP1–DP5), we developed **MAVIS (Multi-Agent Virtual Interactive Synergy)**, a multi-agent system for collaborative preference-based planning in virtual workspace (Figure 1). This section first introduces MAVIS's overall architecture, then walks through a representative interaction flow example, and finally details its core components, Taskformer and CoNavigator, together with their technical implementation.

4.1 System Design Overview

To address the dual challenges of structuring multi-agent-human collaboration and presenting complex information in an accessible manner, MAVIS is architected as a two-part system.

The **Taskformer** component (Figure 2, Figure 4.a) operates as the backend. Rather than relying on rigid domain specifications [1] or returning static plans [52], Taskformer introduces an incremental, user-centered collaboration mechanism. It identifies key planning dimensions from brief task descriptions, recommends specialized expert agents, and generates a staged guideline that determines how agents enter the interaction. It then orchestrates collaboration so that each agent engages the user through focused, proactive dialogue, progressively expanding the discussion across dimensions while maintaining depth. Conflicts and trade-offs are resolved through transparent multi-agent negotiation, with users able to intervene or redirect at any point.

Complementing this, **CoNavigator** (Figure 3, Figure 4.b) serves as the interactive frontend. It mitigates cognitive burden by helping users perceive structure, differentiate agent roles, and maintain awareness of the evolving plan. Taskformer's reasoning is externalized into a spatially organized visual environment where embodied agents, step-linked summaries, and natural interaction transform dense textual exchanges into a situated collaborative experience.

Together, Taskformer and CoNavigator form an integrated architecture: Taskformer structures how plans are constructed, while CoNavigator shapes how users perceive and steer the process. The following case example illustrates how these components operate in concert during a complete planning session.

4.2 MAVIS Interaction Flow through a Case Example

To ground the system architecture in use, we present a representative case derived from real user interactions (Figure 1). Consider Tony, a student organization leader, needs to create a promotional video for recruitment. With limited production experience, he finds it hard to anticipate all relevant considerations in advance and easy to lose track of trade-offs when juggling budget, timing, and creative goals. As initial ideas remain fragmented and difficult to reconcile, he turns to MAVIS to structure and refine his plan.

Automatic Role Definition and Guideline Generation (Figure 1.a). Tony first uses a web-based configuration interface (Figure 5) to enter a short description of his task. Instead of requiring detailed prompts, MAVIS analyzes this brief input, identifies relevant expertise areas (e.g., concept development, scriptwriting, cinematography, project management), and, via a two-step LLM pipeline (Section 4.3.1), proposes both a set of expert agents and a staged guideline specifying when each agent will join and what each stage will focus on. Tony reviews and approves this guideline as the roadmap for collaboration.

Guided Entry into Collaboration (Figure 1.b). After confirming the guideline, Tony enters the MAVIS virtual workspace (Figure 3), here using the VR instance, where embodied expert agents are coordinated by a Project Manager. Rather than passively waiting for user instructions or prematurely attempting broad solutions, the Project Manager follows the guideline to initiate the session, first eliciting organizational context and then inviting the Concept Developer to explore the video's theme and style. This establishes a focused starting point anchored in Tony's goals.

Iterative Interaction and Preference Elicitation for Individual Dimensions (Figure 1.c). MAVIS emphasizes depth by addressing each planning dimension through dedicated dialog. So instead of immediately presenting full scripts or shot lists, the Concept Developer engages Tony in a conversational exchange, drawing on his favorite films and past events to uncover implicit expectations. This iterative process surfaces hidden preferences and refines candidate frameworks before moving forward.

Incremental Expansion and Cross-Agent Collaboration (Figure 1.d). Once the conceptual direction is sufficiently clarified, MAVIS incrementally expands the discussion by introducing additional agents according to the guideline. The Concept Developer collaborates with the Scriptwriter to translate themes into a coherent storyline, while other agents observe or join when their expertise becomes relevant. Information is shared across agents, who negotiate trade-offs and update intermediate outputs. Tony can follow this process, ask questions, or override suggestions, benefiting from multiple perspectives without facing an unstructured multi-agent debate.

Constraint Optimization and Cascading Adjustments (Figure 1.e): A key capability of MAVIS is managing interdependent constraints in real time. Rather than requiring users to manually revise affected elements when conflicts arise, agents proactively optimize trade-offs while keeping the process transparent. For example, given budget limitations, the Cinematographer and Project Manager collaboratively revise costly elements such as drone footage. When Tony later requests an additional interview segment, MAVIS automatically propagates updates across all dependent tasks, ensuring feasibility without burdening the user with recursive adjustments

After iterating through these stages, Tony obtains a comprehensive, preference-aligned production plan that he can export as documentation (Figure 1.f).

4.3 Taskformer: An Incremental Collaboration Mechanism for Preference-based Planning

To delve deeper into the mechanisms behind the walkthrough, this section introduces Taskformer, the backend that structures and coordinates incremental multi-agent–human collaboration. Taskformer formalizes MAVIS's staged planning process through five synergistic stages—Guideline Generation, Incremental Expansion, Focused Exploration, Collaborative Negotiation, and Consolidation (Figure 2)—which together ensure both comprehensive coverage of planning dimensions and alignment with user preferences. These stages directly operationalize design principles DP1–DP4.

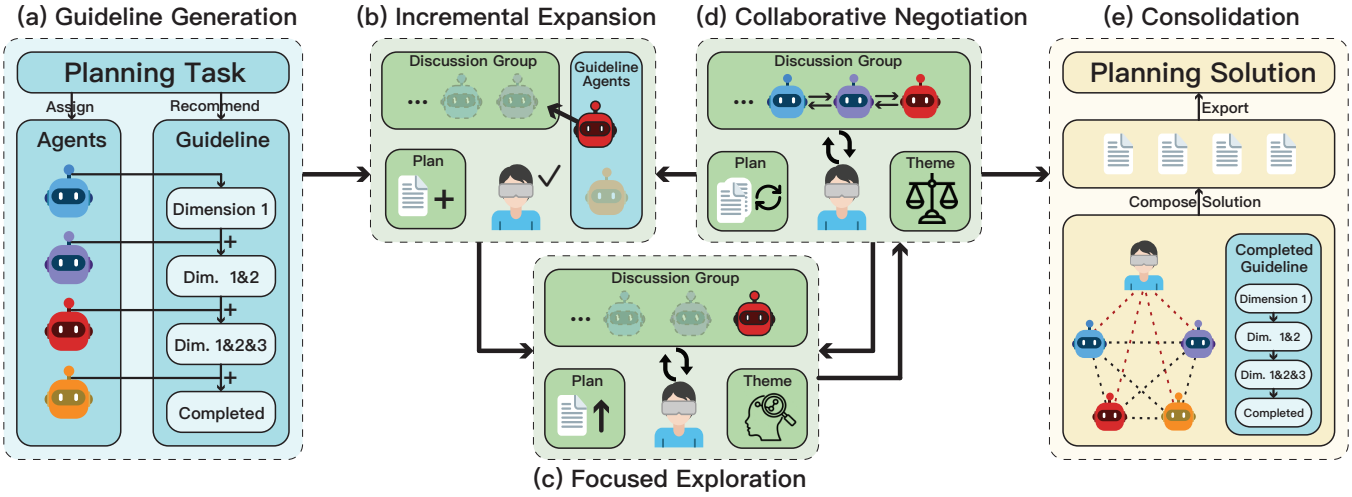


Figure 2: Taskformer Mechanism Architecture, depicting five key operational stages: (a) Guideline Generation, (b) Incremental Expansion, (c) Focused Exploration, (d) Collaborative Negotiation, and (e) Consolidation.

4.3.1 Guideline Generation Stage. The Guideline Generation Stage (Figure 2.a) initiates Taskformer by transforming a user's brief task description into a structured roadmap for multi-agent-human collaboration. Rather than requiring users to provide exhaustive specifications, the system accepts short and informal inputs (e.g., "plan a science club recruitment video"), and then decomposes them into key dimensions and constraints such as budget, timeline, or content requirements. Based on this analysis, Taskformer automatically recommends a set of specialized expert agents and defines their responsibilities. It also generates a staged guideline—a sequenced plan specifying when each agent should join the discussion and which aspects should be prioritized at each stage.

This process is implemented through two consecutive LLM queries (see Appendix A for details). The first query analyzes the description of the task provided by the user, extracting constraints, temporal requirements, and domain-specific considerations. Based on this analysis, the LLM recommends a set of expert agents together with their designated responsibilities. Once this step succeeds, the second query decomposes the task into a logical, incremental sequence of stages, moving from basic to more advanced considerations. Each stage specifies which agent should join and what thematic focus should be addressed, thereby producing a structured guideline that outlines both role assignments and stage-wise progression. Together, the recommended agents and the staged guideline establish a reproducible blueprint for the planning process. Before proceeding, the system invites the user to review and, if necessary, revise this output, ensuring transparency and alignment with their intent.

This stage operationalizes **DP1**. Since complex tasks involve interdependent dimensions that are difficult to anticipate in one step, the staged guideline and explicit role definitions allow users to approach planning progressively rather than attempting to enumerate all requirements upfront.

4.3.2 Incremental Expansion Stage. The Incremental Expansion Stage (Figure 2.b) represents the core of Taskformer's incremental collaboration design. Its role is to guide the user along the staged guideline and introduce new agents once the current dimension has been sufficiently explored. This design contrasts with conventional multi-agent systems, where users are often confronted with a large set of pre-defined agents without clear differentiation of roles or interaction order. As a result, users may either engage with a single agent superficially or become overwhelmed by unstructured interactions, losing awareness of overall task progress. Incremental Expansion ensures that users receive clear, stepwise guidance, while maintaining flexibility for experienced users to deviate, revisit, or extend stages as needed.

This stage is implemented by the designated *Project Manager* agent (see Appendix A for details) and can be triggered from the Guideline Generation Stage (Section 4.3.1) or later Collaborative Negotiation (Section 4.3.4). When triggered, the *Project Manager* performs two actions. First, it compares the current progress with the guideline, then provides the user with a concise summary of completed work and the thematic focus of the next stage. The guideline functions as a soft scaffold: users may return to previous stages, retry steps, or add new steps at any point. Second, once the user confirms to proceed, the *Project Manager* recommends the next expert agent according to the task guideline, introduces its role, and transitions the discussion to the Focused Exploration Stage (Section 4.3.3). Agents introduced in this way are dynamically integrated into the discussion, while users retain the option to directly engage with any agent.

By structuring when and how agents appear, this stage instantiates **DP3**'s call for multi-perspective integration with transparent progress. It prevents users from having to manage all dimensions simultaneously, yet keeps the global roadmap visible, reducing cognitive burden for novices while still allowing experts to flexibly reshape the workflow.

4.3.3 Focused Exploration Stage. The Focused Exploration Stage (Figure 2.c) enables proactive, in-depth interactions between users and expert agents. Each newly introduced agent focuses on its expertise and engages the user through iterative conversations. Rather than presenting immediate solutions, agents begin with approachable questions and gradually progress toward task-specific inquiries. This staged dialogue uncovers implicit preferences, fills knowledge gaps, and ensures that each dimension is explored with sufficient depth before the process expands further.

The stage is implemented through two mechanisms: role definition and user profiling. For role definition, each agent is initialized with structured behavioral instructions (see Appendix A for details). These ensure that (1) agents must collect adequate contextual information before making recommendations; (2) conversations should begin with general, low-stakes questions and progress toward specialized inquiries, preventing users from feeling overwhelmed; and (3) when offering decisions, agents must provide clear, well-structured options that help users weigh alternatives. These instructions ensure that all agents maintain a proactive and iterative interaction style.

The second mechanism, user profiling, maintains continuity across multiple interaction cycles. As conversations accumulate, the system records user inputs into a shared log. Once the data reaches a predefined threshold (a tunable parameter balancing API cost and feedback timeliness; empirically set to every three user turns based on pilot tuning), an LLM is invoked to analyze recent exchanges alongside the existing user profile (see Appendix A for details). The output updates the profile in natural language, stored in a shared memory accessible to all agents. This process keeps agents context-aware, avoids redundant questioning, and allows them to build on evolving preferences.

This stage is designed to fulfill **DP2**. Since users often have limited domain expertise when solving complex problems, they tend to rely on ad-hoc assumptions when working independently. Existing LLM systems are predominantly reactive, waiting for user instructions, which can lead to vague or insufficient inputs and unsatisfactory results. By contrast, this stage places users in a "served" state, where agents proactively anticipate information needs, surface hidden preferences, and provide structured guidance while gradually enriching users' domain understanding.

4.3.4 Collaborative Negotiation Stage. The Collaborative Negotiation Stage (Figure 2.d) coordinates interactions among multiple agents and between agents and the user, once a task involves interdependent dimensions or requires balancing competing constraints. Unlike conventional systems, where users must manually manage trade-offs or sequentially adjust each affected component, Taskformer enables agents to negotiate across perspectives while keeping users actively involved. This design prevents the "black-box" drift common in multi-agent systems, where agents evolve their own discussions beyond user control, by ensuring transparency and intervention opportunities throughout the process.

This stage provides three main advantages. First, it supports multi-dimensional trade-off resolution: when conflicting requirements such as budget, timeline, or quality arise, agents contribute their domain-specific perspectives and iteratively negotiate adjustments, yielding plans that better reflect user preferences. Second, it ensures transparent multi-agent-human collaboration and user control: all communications occur in conversational form, allowing users to observe ongoing coordination, intervene selectively, or redirect discussions at any time. Agents share the burden of balancing trade-offs while keeping the negotiation process observable and steerable, balancing system autonomy and user agency. Third, the stage enables cascading adjustments: when a user modification affects multiple dimensions, agents automatically propagate changes across relevant tasks, sparing the user from repetitive manual edits.

Implementation relies on a custom agent selection mechanism. According to the role definitions established in earlier stages, each agent is provided with descriptions of other agents' roles, and instructed to output structured messages with two fields: "target" and "answer". The "target" selection is determined according to four rules: (1) request information from the user when needed, (2) consult a specialized agent if expertise beyond its role is required, (3) coordinate iteratively with other agents until a resolution is reached, and (4) when user modifications alter task parameters, involve all affected agents sequentially to ensure consistent updates. In addition, whenever a user communicates with or interrupts a specific agent, the system temporarily halts the current message flow and stores the existing conversation history. Once the user completes this ad-hoc intervention, the system re-initiates the calling flow for that agent, enriched with the saved history and updated user information to maintain continuity. Together, these mechanisms formalize inter-agent coordination while preserving transparency and allowing users to intervene seamlessly without disrupting the overall process.

4.3.5 Consolidation Stage. The *Consolidation Stage* (Figure 2.e) aggregates and finalizes the collaboratively produced plan for export and review. During interaction, the system detects agent turns that contain substantive solution content, incrementally merges them into a structured task solution via an LLM, and at the end of the session exports the refined plan in Markdown format (see Appendix A for details).

4.4 CoNavigator: Spatial Visualization for Multi-Agent-Human Collaboration in Virtual Workspace

CoNavigator is MAVIS's interactive frontend, responsible for rendering Taskformer's internal processes into an intelligible collaborative environment (Figure 3). While Taskformer structures multi-agent reasoning and negotiation, its operation inevitably generates extensive textual content, ranging from inter-agent exchanges to evolving plans and supporting context. In conventional LLM and multi-agent systems, this information often remains confined to linear logs, making it difficult for users to digest, distinguish among agents, or maintain awareness of progress. Such overload obscures the advantages of multi-agent–human collaboration and motivates **DP5**: reducing cognitive burden through visualization.

To address this, MAVIS introduces **CoNavigator**, a spatial visualization mechanism that externalizes agent reasoning and distributes information across space. By organizing content spatially and embodying agents as interactive avatars, CoNavigator helps users

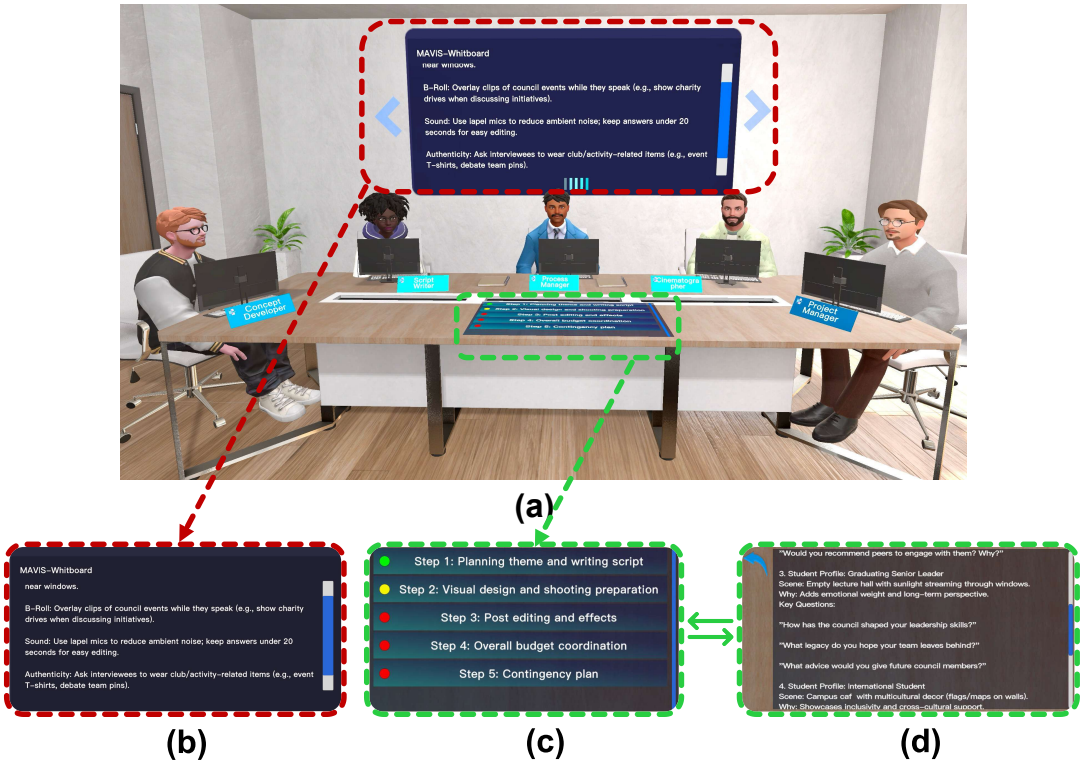


Figure 3: The virtual workspace implemented using CoNavigator, comprising (a) embodied agents within a realistic virtual workspace, (b) a shared interactive whiteboard, (c) a progress panel that transitions to (d) summarized planning histories when the user taps on specific step names. A back button in (d) allows users to return to the real-time progress view.

follow the incremental planning process, distinguish perspectives, and engage with agents in a socially grounded manner.

CoNavigator comprises two complementary modules. The *Virtual Space and Information Visualization Module* spatially organizes planning content into contextual summaries, progress indicators, and shared boards, allowing users to access complex information without being overwhelmed. The *Embodied Agents and Intuitive Interaction Module* presents agents as anthropomorphic avatars that communicate through voice and modality-appropriate input, making roles and negotiation processes easier to follow. Together, these modules operationalize **DP5** by making multi-agent collaboration more transparent and cognitively manageable, enabling users to focus on steering the plan rather than parsing raw text.

4.4.1 Virtual Space and Information Visualization Module. The Virtual Space and Information Visualization Module constructs the collaborative workspace and organizes the extensive textual content generated during multi-agent planning into accessible, structured views. Its goal is to enhance usability, reduce cognitive burden, and sustain creativity by spatially externalizing information that would otherwise remain buried in linear dialogues.

The virtual environment adopts a realistic office setting (Figure 3.a), creating a familiar context that promotes collaborative engagement and original thinking [28]. Within this environment, users and embodied agents convene around a virtual meeting table, oriented toward a large interactive whiteboard (Figure 3.b). A progress indicator panel in front of the user (Figure 3.c) displays the staged guideline and current step. Selecting any completed step opens a summarized view of the associated planning history (Figure 3.d), enabling efficient backtracking and reference.

At the core of this module is an information segmentation and summarization pipeline. During discussion, an LLM monitors agent outputs (see Appendix A for details). Whenever an agent produces content exceeding the length threshold (same tunable in Section 4.3.5) or lacking conversational style, the system separates it into two forms: (1) *communication messages*, concise and dialogue-oriented, delivered directly to the user; and (2) *display messages*, longer and content-rich, visualized on the interactive whiteboard (Figure 3.b). Users can scroll through display messages via hand gestures in VR or mouse input on desktop, while historical messages are archived and browsable via pagination, supporting structured retrieval and iterative refinement.

When the accumulation of *display messages* reaches a threshold, a second LLM is triggered to summarize and classify content (see Appendix A for details). The system distinguishes between partial planning results (e.g., intermediate solutions) and contextual information (e.g., background knowledge, rationales), infers current

progress based on planning results, and updates the progress indicator panel accordingly. Each completed step can be expanded to reveal concise summaries, with full histories available within expandable sections (Figure 3.d). This layered design allows users to quickly access both the overall trajectory and step-level details, balancing transparency with cognitive manageability.

4.4.2 Embodied Agents and Intuitive Interaction Module. This module embodies Taskformer's expert agents as avatars within the virtual environment, making their roles, perspectives, and interactions transparent to the user. It addresses limitations of conventional multi-agent systems, where distinguishing agents' contributions is often difficult and text-heavy dialogues hinder sustained engagement. Guided by DP5, we adopt an information-centered design that emphasizes making planning information inspectable and navigable. By presenting agents as socially present, human-like entities and supporting natural exchanges, the module creates an interaction experience closer to real-world group discussions, encouraging active participation while reducing cognitive effort [37].

The design combines two elements: embodied agents and natural interaction techniques. For embodiment, each agent is rendered as a distinctive avatar with a unique appearance and voice to support role attribution. We generate avatars with Ready Player Me[4], which provides stylized, inclusive characters that are commonly used in HCI systems [18, 36] and can mitigate uncanny-valley concerns in VR [62]. To enhance social presence [21, 75], avatar behaviors are context-aware: speech is accompanied by gesture-aligned animations and sustained eye contact. These behaviors are implemented through a state machine driven by Taskformer outputs, with motion-capture animations from Mixamo[5] and speech synchronization via Oculus Lipsync. Spoken dialogue is generated using the VITS text-to-speech model [43], while synchronized text bubbles next to avatars provide visual clarity.

For natural interaction, we maintain an information-centered rationale by using lightweight, modality-appropriate actions that support interpreting, comparing, and revising planning information (e.g., addressing an expert, browsing shared notes), while mitigating the risk of distracting users from comparative reasoning [60]. In VR, users address an agent through head orientation and initiate interaction with a pinch gesture (i.e., look-to-talk), while browsing the whiteboard and history panels with hand-based scrolling and selection. On desktop, interaction is mediated via mouse-click selection and conventional scrolling. User speech is captured and transcribed in real time, with recognized content displayed as immediate feedback next to the addressed agent. Agent outputs are drawn from the *communication messages* described in Section 4.4.1, ensuring concise, dialogue-appropriate responses that are both vocalized and shown in speech bubbles. Users can freely interrupt or redirect interactions, with avatars responding adaptively while maintaining engagement cues such as eye contact.

By combining embodiment with intuitive interaction, this module helps users follow multi-agent exchanges, associate perspectives with distinct roles, and remain actively involved in the planning process. It turns otherwise abstract, text-heavy reasoning into a socially grounded, more tractable collaborative experience.

[4]https://readyplayer.me/
[5]https://www.mixamo.com/

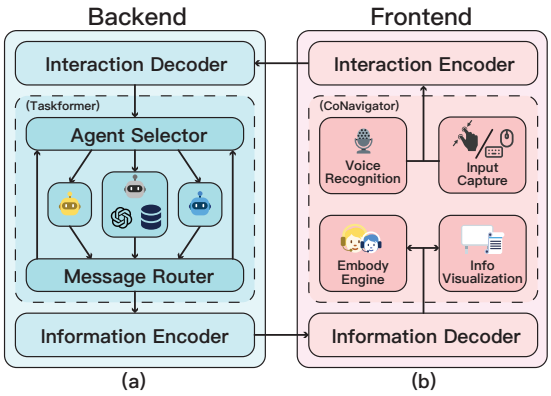


Figure 4: System architecture of MAVIS. (a) The backend (b) The frontend

4.5 System Implementation

Sections 4.3 and 4.4 outlined MAVIS's conceptual design: Taskformer structures incremental multi-agent–human collaboration, while CoNavigator externalizes this process in a virtual workspace. We implement MAVIS as a two-part architecture (Figure 4) with a Python backend for multi-agent coordination and a Unity-based frontend for interaction. The two components communicate over a local WebSocket connection, with round-trip latencies at millisecond-level that were not perceptible to users.

The backend (Figure 4.a) operationalizes Taskformer through a message-driven pipeline. Incoming interaction data from the frontend (recognized speech, target agent) are parsed by an *Interaction Decoder* and routed into Taskformer. Each agent encapsulates an OpenAI API client and a dedicated memory queue, implemented atop AutoGen [81] and LangChain. An *Agent Selector* activates the appropriate agent based on the *target* field in messages or explicit user requests, while a *Message Router* directs outputs either back to the user or to other agents. In parallel, the *Information Encoder* converts intermediate exchanges and partial solutions into structured records consumed by CoNavigator.

Although AutoGen provides infrastructure for agent communication, Taskformer specifies the higher-order logic for incremental, human-centered collaboration (Section 2.2). Accordingly, we use AutoGen only for low-level agent definitions and employ LangChain to re-architect the communication pipeline, message routing, and parsing mechanisms described above. Moreover, since AutoGen conceptually treats users as peripheral input sources[6], we also redesigned the human-in-the-loop logic. Specifically, we incorporated the user as an equal participant alongside other agents and enabled real-time intervention and interruption of ongoing dialogues by dynamically controlling the communication pipeline. This yields a fully connected multi-agent–human collaboration structure, and directly supports Taskformer's five-stage preference-based planning mechanism (as detailed in Section 4.3).

The Unity frontend (Figure 4.b) implements CoNavigator (Section 4.4; Figure 3). We deployed two instances of this frontend to

[6]https://microsoft.github.io/autogen/stable/user-guide/agentchat-user-guide/tutorial/human-in-the-loop.html

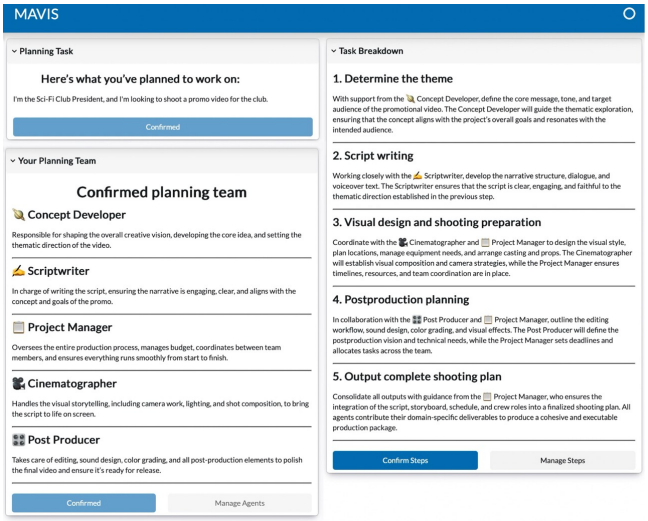


Figure 5: The web-based configuration interface for the Guideline Generation stage. Users provide a brief task description, review and optionally edit the automatically recommended agent roles (left), and then inspect or refine the staged guideline generated from these inputs (right).

support our comparative study: an VR version on the Meta Quest 3 and a desktop version for PC. Messages from the backend are parsed and rendered either as avatar speech (TTS plus bubbles) or as updates to shared boards, progress indicators, and contextual summaries, while user inputs are captured via speech recognition (iFlytek STT API) and platform-specific controls (MetaSDK hand tracking in VR; mouse/keyboard on desktop) and sent back as structured messages.

For the Guideline Generation (Section 4.3.1), users must read and edit relatively dense text. Since in-situ text entry in VR is slower, more error-prone, and more frustrating than desktop typing for text-intensive tasks [7, 13, 38, 77], we perform this step in a lightweight web-based configuration interface built with HoloViz Panel[7] (Figure 5). Users refine the recommended agents and staged guideline on desktop; the configuration is serialized as JSON and then used to launch the low-text, visualization-focused virtual workspace for the main embodied collaboration.

5 Evaluation Studies

To evaluate MAVIS and its design components, we conducted three progressively connected studies (Figure 6). Study 1 quantitatively assessed the effectiveness of the *Taskformer* component in isolation, focusing on planning performance metrics. Study 2 qualitatively examined the impact of the *CoNavigator* visualization component on subjective experiences such as engagement and sensemaking. Study 3 evaluated MAVIS's cross-task versatility and cross-platform generalizability in an expert study, in which participants used both *MAVIS (VR)* and *MAVIS-Desktop* to compare the influence of interface modality on various domain-relevant real-world planning tasks. Together, these studies provide complementary evidence of MAVIS's performance, user experience, and applicability across tasks and interaction modalities.

5.1 Study 1: Quantitative Evaluation of the Taskformer

The goal of Study 1 was to evaluate the effectiveness of the Taskformer mechanism independently of other components. We implemented a simplified multi-agent system with the Taskformer component, referred to as MAS (Multi-Agent Synergy), which exposed only Taskformer's core functionality through a ChatGPT-like web interface designed to minimize UI-related confounds (Figure 7; see Appendix B for design rationale and a side-by-side comparison). MAS was compared with a baseline system (GPT-4o) in a within-subjects design focused on travel planning, a representative preference-based planning task that requires trade-offs across budget, transportation, sightseeing, and dining.

When selecting a baseline for this experiment, we considered several practical and methodological factors. Popular open-source multi-agent frameworks such as AutoGen [81], CrewAI[8], and CAMEL [51] require users to manually define agent roles, goals, and collaboration structures before any interaction can take place. For ordinary end users, such configuration presupposes a level of AI literacy and skill that is not yet widespread [44]. Using such systems as baselines would therefore necessitate extensive instruction or researcher-led configuration, which would obscure the very usability challenges that Taskformer aims to address. Meanwhile, academic multi-agent systems like MetaGPT [34], ChatDev [61], and StorySage [73] focus on specialized domains (e.g., programming or writing) and are best viewed as research prototypes rather than widely adopted, general-purpose planning tools. In contrast, ChatGPT remain the most commonly used tools for real-world planning tasks [10], offering a mature, well-understood, and user-accessible benchmark. For these reasons, we selected GPT-4o as a practical baseline for isolating Taskformer's contribution to preference-based planning performance.

5.1.1 Methods. We recruited 14 participants from university student communities, with one outlier excluded due to substantial deviation from the protocol instructions, resulting in 13 participants in the final analysis (4 female, 9 male; $M_{\text{age}} = 20.9$, $SD = 1.38$). Regarding prior experience, all participants reported using LLMs at least weekly (1–3 times per week), and 84.6% reported using them almost daily. Each session lasted approximately one hour, and participants received 50 CNY as compensation.

The study followed a within-subjects design. Sessions were conducted in a lab room with participants seated throughout. After providing informed consent, participants were asked to recall a close friend and select two potential travel destinations based on both individuals' preferences. A short pre-interview was conducted to collect their expectations and travel-related preferences, forming the baseline for subsequent analysis. Participants then interacted with the first system (MAS or GPT-4o[9], counterbalanced to avoid order effects) via a laptop-based web interface placed on the desk. Each session began with a short familiarization period, after which

[7] https://panel.holoviz.org/

[8] https://crewai.com/

[9] https://chatgpt.com

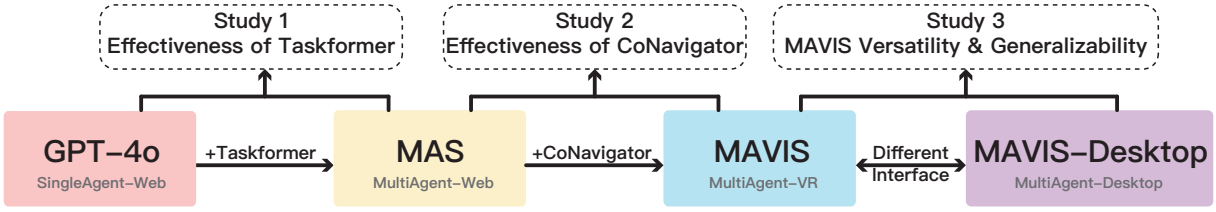


Figure 6: Relationships Among the Three Evaluation Studies.

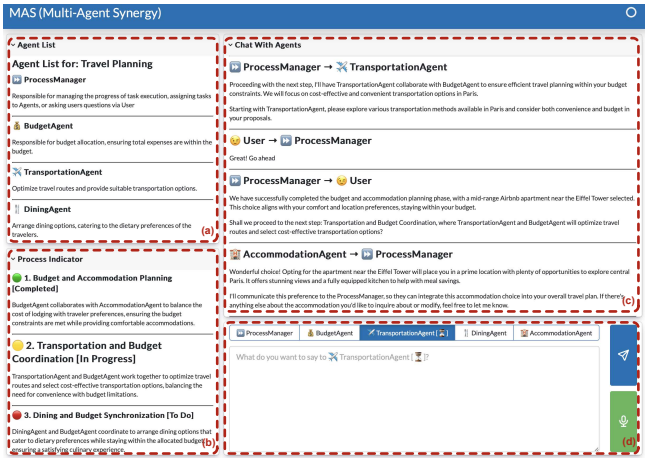


Figure 7: Multi-Agent Synergy (MAS) system Interface. (a) agent lineup panel displaying currently assigned agents' roles, (b) task progression panel indicating subtask flow and overall progress, (c) interaction history panel displaying user–agent exchanges, and (d) input panel for selecting a target agent and submitting voice or text-based queries.

participants used the assigned system to plan a complete itinerary. Once participants were satisfied with the plan, the session concluded. After a 10-minute break, participants repeated the process with the other system for the second destination.

At the end of the study, participants compared outputs from the two systems and noted whether new preferences or expectations had been surfaced during the interaction. All interactions were logged, and interviews and outputs were recorded for later analysis. To determine statistical significance between Taskformer (Condition A) and the baseline (Condition B), we performed a paired t-test for metrics following a normal distribution and a Wilcoxon signed-rank test when this assumption was not met.

5.1.2 Metrics. Our evaluation focused on four metrics, selected to capture both user engagement and the quality of preference elicitation and integration, which are central to preference-based planning.

1. User Message Count. We measured the number of participant-authored submissions sent via the text-input box in each condition. We use this metric to index users' active engagement and willingness to contribute information resulting from the design of Taskformer. Because preference-based planning is highly user-defined and difficult to standardize quantitatively, this metric should be interpreted together with Preference Points and their derivatives (Preference Expressed/Met) below.

2. Preference Points. To establish a ground truth, we collected preferences during the pre-interview. Participants were asked structured questions covering aspects such as budget, transportation, food, and activities. Audio recordings were transcribed and analyzed independently by two researchers using cross-checking methods. The resulting set of Preference Points represented the comprehensive pool of each participant's individual and friend-related preferences.

3. Preference Expressed Rate. During interaction, whenever a participant explicitly mentioned a preference (e.g., "avoid long bus rides"), it was coded as an Expressed Preference Point. We then calculated the proportion of expressed points relative to the total number of Preference Points for that participant, i.e., Preference Expressed Rate $= \frac{\text{Expressed Preference Points}}{\text{Total Preference Points}}$. This metric evaluates whether the system effectively supported users in articulating their known preferences.

4. Preference Met Rate. To evaluate output quality, we analyzed the final travel plans. A preference was marked as "met" if it was reflected in the generated plan. The rate was computed as Preference Met Rate $= \frac{\text{Met Preference Points}}{\text{Total Preference Points}}$, which reflects how well system outputs aligned with user needs and preferences.

Together, these metrics form a pipeline of evaluation: (i) whether users were actively engaged (User Message Count), (ii)whether engagement facilitated preference articulation (Preference Expressed Rate), and (iii) whether articulated preferences were successfully incorporated into outcomes (Preference Met Rate). These metrics support our claim that Taskformer enhances both the process and results of preference-based planning.

5.1.3 Results. The experiment results show that Taskformer notably increased user engagement compared to the baseline (Figure 8.a). Participants using MAS sent an average of 13.15 messages (SD=3.46) versus 6.31 with GPT-4o (SD=2.69; t=7.24, p<0.001, d=2.01). Here, User Message Count captures participant-authored submissions via identical input controls across conditions, so the higher count reflects that Taskformer prompted users to contribute more information, giving them more opportunity to articulate their preferences. This interpretation is further supported by the higher Preference Expressed rates reported below. Moreover, the User Messages sent per minute did not differ significantly (MAS: M=1.11;

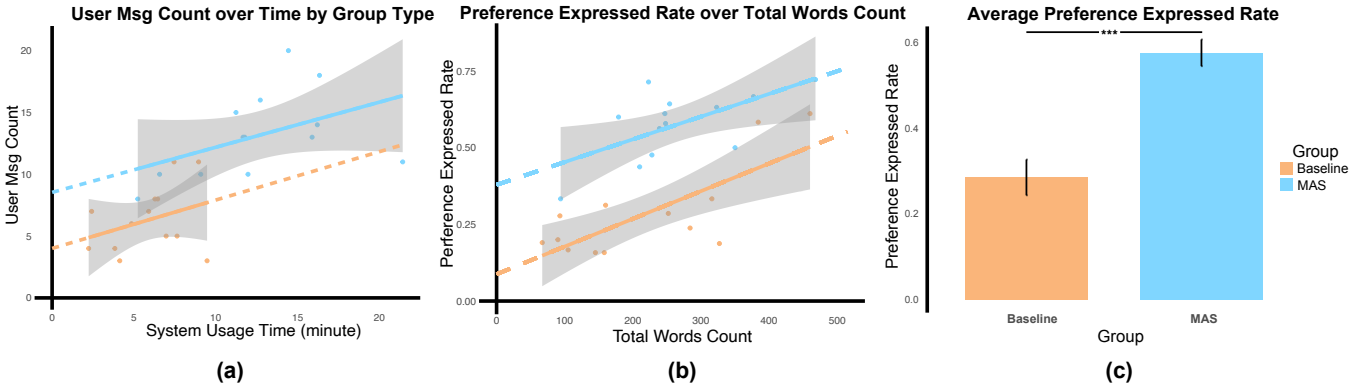


Figure 8: Comparison of (a) user-authored message count over session time, (b) preference expression rate relative to total word count and (c) average preference expression rates between MAS and GPT-4o groups. The asterisks '*' denote statistical significance: * p<0.05, ** p<0.01, *** p<0.001. This significance notation applies to subsequent figures.

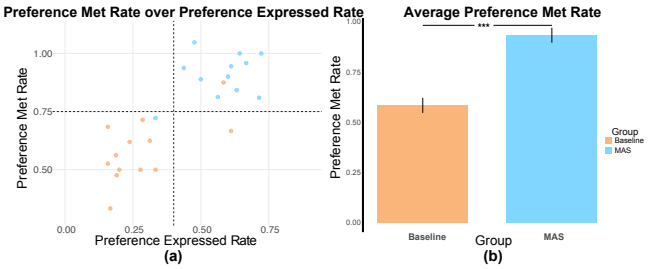


Figure 9: Relationship between expressed preferences and preferences met (a), and comparison of average preference met rates (b) between MAS and GPT-4o groups.

GPT-4o: M=1.22; t=-0.58, p=0.57), indicating richer engagement without additional interactional overhead.

Beyond engagement, Taskformer facilitated more effective preference articulation. A positive correlation was observed between the amount of text input and the rate of expressed preferences (Figure 8.b; r=0.64, p<0.001). Although total number of user-input words is comparable between systems (MAS: M = 264.62 words, SD = 95.11; GPT-4o: M = 218.54, SD = 126.82; t=1.95, p=0.075, d=0.541), MAS elicited a significantly higher Preference Expressed Rate (Figure 8.c; MAS: M = 0.58, SD = 0.11; GPT-4o: M = 0.29, SD = 0.15; t=6.66, p<0.001, d=1.85). This suggests that Taskformer's staged and proactive dialogues enabled participants to surface implicit constraints and preferences that would otherwise remain unarticulated.

These improvements in articulation translated into higher-quality outputs. A strong positive correlation was found between the rate of preferences expressed and the rate of preferences met in the final plans (Figure 9.a; r=0.78, p<0.001). Taskformer significantly outperformed the baseline in this respect, with a higher Preference Met Rate (Figure 9.b; MAS: M = 0.93, SD = 0.14; GPT-4o: M = 0.58, SD = 0.14; t=8.71, p<0.001, d=2.42). This demonstrates that Taskformer not only encouraged more active participation but also guided users toward outputs that more reflect their individual preferences, validating its role in optimizing preference-based planning.

To examine whether gender moderated Taskformer's within-subject gains, we computed per-participant deltas ($\Delta = MAS - GPT$) for the three primary outcomes (User Message Count, Preference Expressed Rate, and Preference Met Rate). As Shapiro–Wilk tests suggested approximate normality, we compared genders using Welch's t-tests. In the originally completed sample (N = 13; 9 male, 4 female), we found no evidence of gender association on the deltas for UserMessageCount ($t = -0.34, p = 0.741, d = -0.16$), ExpectExpressRate ($t = -0.67, p = 0.526, d = -0.40$), or ExpectMetRate ($t = -0.60, p = 0.566, d = -0.24$). Consistently, the paired improvements of MAS over GPT remained significant within both gender subgroups (paired t-tests: UserMessageCount, male t = 5.08, p = 0.001, female t = 6.54, p = 0.007; ExpectExpressRate, male t = 5.01, p = 0.001, female t = 4.23, p = 0.024; ExpectMetRate, male t = 5.80, $p < 0.001$, female t = 25.67, $p < 0.001$). We report the main results based on the original sample; as supplementary analysis, we later recruited four additional female participants (M_{age} = 21.75, SD = 0.83) using the same protocol and apparatus, and re-ran the same checks on the expanded dataset, which yielded the same pattern of non-significant gender–delta associations (UserMessageCount t = 0.03, p = 0.980; ExpectExpressRate $t = -1.59$, p = 0.134; ExpectMetRate $t = -0.45$, p = 0.661) and significant paired effects within each gender.

5.1.4 Summary. Study 1 provides quantitative evidence that Taskformer's incremental, multi-agent design directly benefits preference-based planning. By structuring interactions into stagewise dialogues and incrementally introducing expert perspectives, Taskformer encouraged users to engage more actively and articulate a higher proportion of their preferences. These articulated preferences were then more consistently reflected in the final outputs, resulting in plans that were significantly better aligned with user needs than those generated by baseline. In short, Taskformer validated its design goals: it improved engagement without increasing effort, surfaced implicit preferences, and translated them into higher-quality, user-centered planning results.

5.2 Study 2: Qualitative Analysis of the CoNavigator's User Experience

Building on Study 1, which demonstrated the effectiveness of Taskformer, Study 2 examined how the CoNavigator visualization component shapes user experience. Given the subjective nature of visualization effects such as engagement, sensemaking, and perceived effectiveness, we adopted a qualitative approach. Participants interacted with both MAVIS (Figure 3, including CoNavigator) and a non-spatial multi-agent baseline (MAS, Figure 7, excluding CoNavigator) to compare differences in engagement, decision-making, and perceived effectiveness during travel planning.

5.2.1 Methods. We recruited 16 participants (5 females, 11 males; $M_{\text{age}} = 21.5$, $SD = 1.41$) from a university community. Regarding prior experience, all participants reported using LLMs at least weekly (1–3 times per week), and 75% reported using them almost daily, while 56.25% had prior VR experience. Due to participant availability during recruitment, the sample was not gender-balanced. However, our analysis follows reflexive thematic analysis, where adequacy is primarily judged by the fit between the dataset and the research questions [9]. Our gender distribution is also comparable to prior work in similar contexts [47, 86]. We therefore deem the dataset adequate for our qualitative aims, while acknowledging this demographic skew as a limitation.

After informed consent, participants planned two distinct travel itineraries, one with MAVIS and one with MAS, similar to Study 1. System order was counterbalanced, and different destinations were assigned to reduce learning effects. Sessions took place in a lab room (Figure 10). MAS was accessed via a laptop-based web interface. For MAVIS, participants first completed Guideline Generation on the same laptop, then wore a Meta Quest 3 headset and used the VR interface via wired PC streaming (4K at 90 Hz) from a Unity 2022.3 application running on a laptop with an NVIDIA RTX 3080 GPU. IPD and strap fit were adjusted before VR use, and participants could pause or remove the headset at any time.

Throughout the session, a think-aloud protocol was used to capture participants' real-time reflections during interaction. After completing both systems, participants joined a semi-structured interview comparing MAVIS, MAS, and their usual planning practices. All sessions were audio-recorded, transcribed, and analyzed using Braun and Clarke's thematic analysis [8]. Two researchers independently coded the data and iteratively refined the codebook. We also collected interaction logs as contextual references to support interpretation of qualitative themes.

5.2.2 Results. The qualitative analysis of the CoNavigator revealed two central themes that characterize its influence on user experience, particularly in how users manage complex information and engage with embodied multi-agent collaboration.

Structured information mechanisms support detailed planning without losing global context. All Participants reported that CoNavigator's information layout made complex content more digestible and helped them balance local details with the overall plan, in contrast to the linear, scroll-based MAS interface.

Theme 1: Segmented and spatialized information reduces overload and supports on-demand referencing. Participants (11/16) emphasized that separating concise verbal queries from detailed contextual

Figure 10: Study 2 apparatus and setup. Right: web interface (MAS) or Guideline Generation page for MAVIS. Left: MAVIS with Meta Quest 3 via wired PC streaming. A facilitator monitored a preview; audio and screen outputs were recorded.

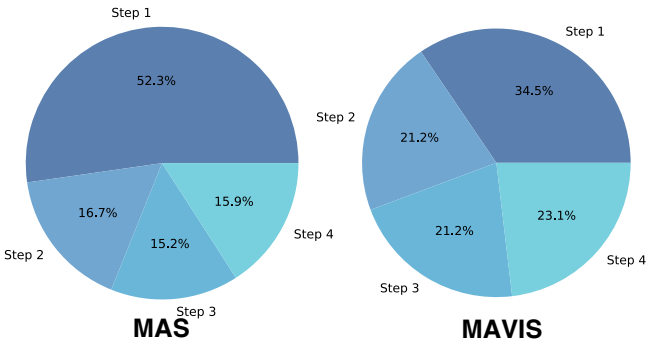


Figure 11: Distribution of planning time across the guideline steps for MAS and MAVIS

information substantially lowered the effort needed to follow the conversation. Agents posed short questions, while the whiteboard displayed richer context. P10 and P14 found that this division made content easier to understand. Many participants (9/16) described treating the whiteboard and summaries as external memory rather than trying to remember everything. P1 noted that hearing an explanation and then seeing it on the board made it easier to refer back to context and reinforced the information twice, making the content more memorable.

Theme 2: The progress panel scaffolds orientation, retrieval, and a sense of progress. The progress indicator panel served as more than a status display. Participants (9/16) used it to reorient when conversations became detail-heavy, to quickly retrieve earlier decisions, and to guide what to do next. P13 noted that it *"helped me keep track of the overall progress when I got stuck in details."* Participants (5/16) reported revisiting completed steps via the panel when they needed to recall earlier agreements. For less experienced planners, the panel acted as a behavioral script; P14 likened completing items on the panel to playing a game, which made the planning process feel more structured and less stressful. Log data aligns with these accounts: MAS users spent over half of their time in the initial guideline step (Step 1: 52.3%), whereas MAVIS users distributed

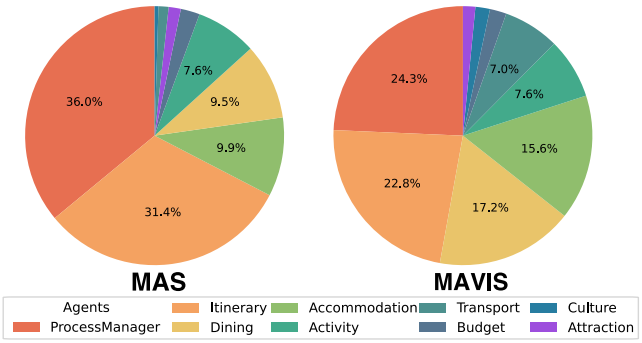


Figure 12: Distribution of interaction frequency with different agents for MAS and MAVIS

their time more evenly across all four steps (Step 1: 34.5%, Steps 2–4: ≈21% each; Figure 11).

Embodiment and voice interaction reshape expression, multi-agent understanding, and social experience. Embodied agents and spoken dialogue encouraged more multi-perspective, and socially grounded interactions than the text-based baseline, although heightened social presence was not uniformly positive.

Theme 3: Anthropomorphism and speech lower inhibition and redistribute attention across specialized agents. Participants (13/16) reported that avatars and voice interaction made the system feel more like a real conversation than form filling, which encouraged richer expression. P12 noted that speaking "felt like explaining things out loud," so they shared more concrete details without over-polishing wording. Interaction logs reflected this shift: while MAS conversations were dominated by the ProcessManager and ItineraryAgent (36.0% and 31.4% of user turns), MAVIS interactions were more evenly distributed across diverse agents (e.g., ItineraryAgent 22.8%, DiningAgent 17.2%, AccommodationAgent 15.6%; Figure 12). Participants explained that distinct avatars and roles helped them "*think clearly and logically from different dimensions*" (P8) and made it natural to address specific agents when switching topics, rather than treating the system as a single monolithic assistant.

Theme 4: Visible multi-agent negotiation and a professional atmosphere scaffold understanding and engagement. Embodiment also shaped how participants understood the planning process. Participants (10/16) described the VR scene as a "*real meeting*" where expert agents visibly negotiated trade-offs. P5 characterized it as experts facing one another to balance different aspects, visualizing the thinking process, which reassured them that constraints were being actively managed. Watching this process sometimes sparked ideas for their own planning; P8 reported that seeing agents adjust itineraries based on dining needs helped them understand how to organize the trip more holistically. The spatial arrangement of agents around the user contributed to a sense of centrality and professionalism. P12 described the experience as an "*exclusive, proactive service where my opinions were respected and recognized*," and P2 felt that the formal collaborative atmosphere prompted clearer, more logical thinking, encouraging more careful engagement than with the web interface.

Theme 5: Social presence motivates some users but introduces performance pressure for others. Social presence, however, functioned as a double-edged sword. Most participants (14/16) reported feeling energized and more accountable under the attention of the avatars. For a small subset (1/16), the same presence was a source of stress. P8 noted that being looked at made them tended to rehearse or structure their wording before speaking. This divergence indicates that while embodied, socially present agents can be a powerful motivator, they may also shift the interaction from spontaneous dialogue toward perceived performance for some users, highlighting the need for adaptable levels of social intensity in future systems.

5.3 Study 3: Expert Evaluation of MAVIS's Versatility and Cross-Platform Generalizability

While Studies 1 and 2 examined component-level effectiveness, Study 3 aims to evaluate MAVIS across real-world domains and interaction modalities. We conducted a within-subjects qualitative study with domain experts in various real-world planning scenarios (Table 1). Each expert used both **MAVIS (VR)** and **MAVIS-Desktop**, a desktop application that implements the same CoNavigator visualization in a non-VR 3D workspace operated via mouse and keyboard, point-and-click selection, and press-to-talk speech input (Figure 13). Using preference-based planning tasks grounded in their own practice, we probed (1) how well MAVIS supports diverse real-world scenarios and (2) how interface modality shapes perceived usability, engagement, and suitability for everyday use.

Figure 13: Study 3 Apparatus. Upper: The experimental setup where a participant interacts with MAVIS-Desktop using a laptop. Lower: A screenshot of the MAVIS-Desktop interface.

5.3.1 Methods. We recruited 14 expert participants (7 males, 7 females; M_{age} = 21.29, SD = 1.83), each with substantial experience in their domain (see Appendix C for participant profiles). Regarding prior experience, all participants reported using LLMs at least weekly (1–3 times per week), and 78.6% reported using them almost daily, while 64.3% had prior VR experience.

Table 1: Task features and characteristics for the expert evaluation in Study 3.

Task / Feature	Collaborative	Decision	Optimize	Coordinative	Interdisciplinary	Creative
Video Production (E01–E04)	✓	×	×	✓	✓	✓
Debate Preparation (E05–E06)	×	✓	×	×	✓	✓
Competition Planning (E07–E08)	✓	✓	✓	✓	×	×
Group Event Planning (E09–E11)	✓	✓	✓	✓	×	✓
Performance Planning (E12–E14)	✓	✓	×	✓	×	✓

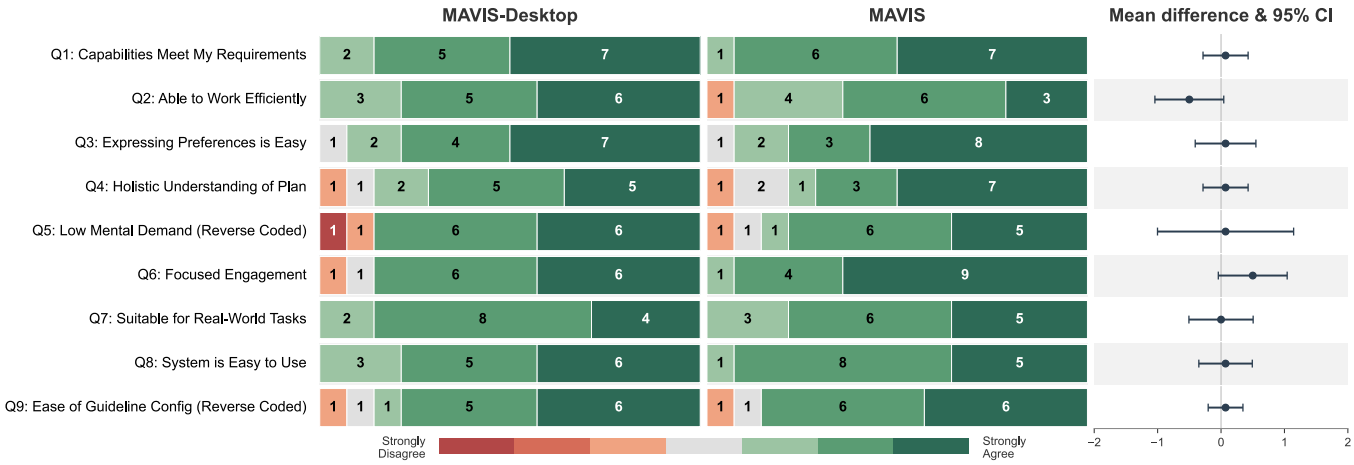


Figure 14: Participants' responses to the 7-point Likert questionnaire comparing MAVIS-Desktop and MAVIS. Stacked bars show response distributions for each question. Dots on the right depict the mean difference (MAVIS – MAVIS-Desktop), with bars indicating 95% confidence intervals.

Each participant completed one domain-relevant planning task twice, once with MAVIS and once with MAVIS-Desktop (see Appendix C for task descriptions). Tasks required balancing multiple constraints and preferences and collectively covered the planning dimensions summarized in Table 1. System order was counterbalanced, and different topics within the same task type (e.g., different debate motions) were used to reduce learning effects. Each session lasted approximately 1.5 hours, and participants received 80 CNY.

The procedure followed the same protocol as Study 2. After informed consent, participants performed the two conditions with think-aloud, and after each condition completed a 7-point Likert questionnaire (Figure 14) comprising seven custom items and two UMUX-Lite items [50] (Q1, Q8). Sessions were conducted with the same setup as Study 2, and were audio-recorded, transcribed, and analyzed using Braun and Clarke's thematic analysis [8], yielding 74 codes grouped into 16 sub-themes and 6 main themes. For questionnaire data, we used paired t-tests for normally distributed metrics and Wilcoxon signed-rank tests otherwise.

5.3.2 Results. Overall, domain experts confirmed both MAVIS (VR) and MAVIS-Desktop as highly capable for their professional tasks (Q1-Capability: MAVIS M=6.43, SD=0.65; MAVIS-Desktop M=6.36, SD=0.74; W=6, p=0.65, r=0.08; Figure 14) and easy to use (Q8-Usability: MAVIS M=6.29, SD=0.61; MAVIS-Desktop M=6.21, SD=0.8; W=12, p=0.71, r=0.06). Both were judged suitable for real-world deployment (Q7-Suitability: MAVIS M=6.14, SD=0.77; MAVIS-Desktop M=6.14, SD=0.66; W=14, p=1, r=0). To further illustrate how MAVIS supports real-world planning scenarios, we selected a representative case (E12) and provided an annotated interaction log in Appendix D. While both the VR and Desktop interfaces successfully supported these complex workflows (UMUX-Lite: MAVIS M=80.94, SD=5.79; MAVIS-Desktop M=80.16, SD=6.62; t=0.563, p=0.583, d=0.15), qualitative feedback revealed modality-dependent differences in how users perceived agent collaboration and managed cognitive effort.

MAVIS generalizes across domains by aligning task complexity and reducing operational costs. Experts in all four domains reported that both interfaces closely matched their real workflows, reduced planning and adjustment effort, and yielded comprehensive, often creatively enriched outputs.

Theme 1: Replicating task complexity and holistic coverage. All Experts reported that the system captured the key dimensions and inter-dependencies of their tasks, even estimating a "*95% restoration*" of real-world task structures (E07). They attributed this to the guideline generation, which surfaced domain-specific dimensions often missed by generic LLM tools (11/14). In debate preparation, E05 felt that "*every dimension we value, from definitions to rebuttals was covered*", and E06 noted that the adversarial agent reproduced

the crucial step of simulating an opponent's attack. In video production, E01 highlighted that "*coordinating schedules across multiple actors and locations was highly usable*," indicating that complex constraints were handled in ways consistent with professional practice.

Theme 2: Reducing operational cost through meta-planning and low-cost adjustment. Experts (9/14) described a shift from manual execution toward meta-planning, where they provided high-level requirements and let agents carry out detailed work. E13 described this shift as "*moving away from sequential information gathering to providing macro-level requirements, allowing parallel execution by agents and saving massive amounts of energy*," consistent with high efficiency ratings (Q2: MAVIS M=5.71, SD=1.07; MAVIS-Desktop M=6.21, SD=0.8; W=2, p=0.068, r=0.34). CoNavigator's context panels further reduced information load; E10 noted that "*dynamically updated context reduced the burden of digging through materials*." Multi-agent coordination also automated cascading adjustments. E03 observed that budget overruns, which normally require massive rework, were instead handled when "*creative and resource agents automatically and jointly adjusted plans to fit new constraints*."

Theme 3: Producing comprehensive and creatively enriched plans. Experts (10/14) described the final outputs as complete and directly usable, with minimal need for post-hoc editing. E04 reported that they "*basically did not spend extra effort adding content*," and that answering agents' questions produced a result they could directly use. Agent collaboration helped catch oversights: E11 highlighted automatically added backup plans for bad weather, and E07 valued "*emergency response measures*" such as backup staff. Beyond completeness, experts appreciated creative support. E03 felt that "*diverse perspectives broke fixed thinking loops*," and E02 was struck by suggestions for strengthening emotional connection in a promotional video, a creative angle they had not originally considered.

Interface modality mediates the experience of embodied multi-agent collaboration. Although VR and Desktop received similarly high capability and usability ratings, experts reported distinct patterns of attention, expression, and social experience: VR amplified immersion and conversational engagement with agents, whereas Desktop favored familiar, text-centric interaction for routine use.

Theme 4: Desktop attenuates embodiment and shifts attention back to text. On Desktop, several experts (6/14) described a mismatch between seeing avatars and interacting via mouse and keyboard. E14 found this disjointed relative to the interaction implied by the avatars, leading to "*a sense of disparity that reduced willingness to interact with them socially*". E05 and E06 characterized Desktop agents as "*traditional AI tools with cartoons*" or "*flat tools*," noting that their attention shifted mainly to text bubbles and panels. This also changed expression style: E06 reported that VR encouraged low-effort "*conversational expression*," whereas the Desktop interface prompted more formal "*written expression*," which they experienced as more cognitively demanding.

Theme 5: VR strengthens immersion, focus, and conversational expression. In VR, experts (11/14) highlighted intuitive, socially grounded interaction and stronger focus. E01 and E12 emphasized that initiating dialogue by simply looking at an expert felt more natural than clicking, reducing friction when switching between agents. E09 noted that wearing a headset reduced visual distractions, helping them devote attention to the task. Even though participants knew agents were not real, E13 described a "*subconscious suggestion of social presence*" in VR that led them to think more carefully before speaking. E03 added that VR "*made the task more interesting and less stressful*," particularly for obligatory tasks.

Theme 6: Desktop offers comfort and familiarity for routine or time-constrained use. Experts (7/14) emphasized the pragmatic advantages of the Desktop interface. Three participants mentioned the headset's weight and prolonged use as limiting factors, preferring the "*ingrained familiarity of Desktop for quick, transactional interactions*" (E08, E11). Despite these physical differences, both systems were rated similarly on efficiency (Q2), and E01 argued that VR's learning curve is temporary, with friction decreasing as users gain experience. Taken together, the modality choice depends on the target scenario: Desktop is optimal for general, low-stakes planning (e.g., weekend trips) where efficiency is more important, whereas VR is uniquely suited for complex, open-ended creative tasks (e.g., video production) where the cognitive gain of social presence and immersion outweighs the physical cost of the headset.

6 General Discussion

6.1 Design Implications

The insights derived from our requirement elicitation study and the subsequent design and evaluation of MAVIS offer several transferable implications for designing human-centered interactive systems beyond preference-based planning tasks.

Treat users as meta-planners guided by staged, role-based collaboration. The requirements study and evaluations show that users rarely know all constraints or preferences upfront, yet they can readily recognize reasonable decompositions and trade-offs once these are made explicit. This suggests that future planning systems should cast users as meta-planners who supervise structure and priorities, rather than as prompt authors responsible for full specifications. For example, in a collaborative design or data-analysis tool, the system can propose an editable roadmap that assigns subtasks to specialized agents, allowing users to focus on high-level thinking while delegating execution. Through meta-operations such as reviewing intermediate results, reordering or skipping stages, testing alternatives, and triggering cascading updates, users can explore ideas faster without losing overall control.

Externalize multi-agent reasoning as layered, navigable structure rather than linear text. Across studies, users struggled with long transcripts but benefited from spatially and structurally separated views of the process. More generally, multi-agent systems should decouple conversational turns from persistent context, maintaining explicit representations of stages, evolving summaries, and constraint state that can be navigated by step instead of by scroll position. Even in non-immersive settings, separate surfaces for dialogue, summaries, and trade-off views can make optimization processes more legible than raw textual explanations. Transparency thus becomes an interaction and information-architecture problem: users need to see how agents respond to changes and how their proposals converge, not only why a final answer is "reasonable."

Use embodiment and modality as levers to shape expression and use cases. Our expert study indicates that the same multi-agent mechanisms are experienced differently in immersive and desktop settings. Embodied agents and situated environments

can encourage narrative, exploratory descriptions and sustained focus, but also impose physical and social costs. Desktop interfaces support rapid, frequent interaction but tend to flatten agents into generic widgets. Rather than asking which modality is better, designers should align modality with task and context. Immersive, socially rich settings are well suited to open-ended, high-impact planning sessions where depth and negotiation matter, while desktop or web interfaces fit routine updates and broad deployment. Hybrid workflows that preserve a consistent multi-agent structure across modalities are a promising direction for future work.

6.2 Limitation and Future Work

While MAVIS demonstrates the potential of spatial multi-agent collaboration for preference-based planning, our findings highlight several limitations for future research.

First, our system relies on a general-purpose LLM, which may under-perform in highly specialized or high-stakes domains. Although experts reported good alignment with their current practice, deeper domain expertise will likely require integrating external knowledge (e.g., retrieval-augmented generation) and, where appropriate, domain-specific fine-tuning. Investigating how staged, multi-agent planning interacts with such extensions remains an open question.

Second, our evaluation treated VR and Desktop as distinct interaction paradigms, yet real-world workflows often fluctuate between creative brainstorming and rigorous execution. As noted in Study 3, VR maximizes cognitive engagement for strategy, while Desktop excels at efficient implementation. The current separation forces users to choose one trade-off over the other. Future work should develop cross-device interfaces that enable smoother transitions, for example, initiating a project in an immersive workspace and migrating outputs to desktop for detailed refinement.

Third, MAVIS currently emphasizes information-centered interaction through avatar-mediated dialogue and stable shared surfaces. Richer VR-specific techniques, such as artifact manipulation, pointing, and object-referenced grounding, could further support coordination and shared understanding. Future work should systematically investigate how these interaction techniques complement spatial information visualization in multi-agent planning, and when their benefits justify the added interaction and implementation complexity.

Finally, Study 3 highlighted social presence as a double-edged sword: motivating for many participants, yet anxiety-inducing for some, especially more introverted users. Our current agents exhibit a uniform social demeanor and fixed seating arrangement. Future systems should explore personality- and context-aware social behaviors, such as modulating gaze, turn-taking intensity, or spatial distance, so that the benefits of embodied collaboration can be preserved while avoiding unnecessary social pressure.

7 Conclusion

We presented MAVIS, a human–AI collaboration system for preference-based planning that combines an incremental multi-agent mechanism (Taskformer) with spatial visualization (CoNavigator). Informed by a requirements elicitation study, MAVIS structures planning as staged collaboration and transparent negotiation while externalizing extensive textual information into spatial visualizations. Across three studies, we showed that this approach improves preference articulation and alignment of plans with user preferences over an LLM baseline, supports more manageable sensemaking through CoNavigator, and generalizes across expert domains and both VR and desktop modalities. These findings yield design implications for future decision-support systems that treat users as meta-planners, expose multi-agent reasoning as an inspectable and steerable process, and select interaction modalities according to task demands.

References

[1] James Allen, Lucian Galescu, Choh Man Teng, and Ian Perera. 2020. Conversational Agents for Complex Collaborative Tasks. *AI Magazine* 41, 4 (2020), 54–78.

[2] Christopher Andrews, Alex Endert, and Chris North. 2010. Space to think: large high-resolution displays for sensemaking. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, Atlanta Georgia USA, 55–64.

[3] Ashraf Bah, Praveen Chandar, and Ben Carterette. 2015. Document Comprehensiveness and User Preferences in Novelty Search Tasks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Santiago Chile, 735–738.

[4] Jorge A. Baier and Sheila A. McIlraith. 2008. Planning with Preferences. *AI Magazine* 29, 4 (2008), 25–36.

[5] Ommi Balamurali, AM Abhishek Sai, Moturi Karthikeya, and Sruthy Anand. 2023. Sentiment analysis for better user experience in tourism chatbot using lstm and llm. In *2023 9th International Conference on Signal Processing and Communication (ICSC)*. IEEE, IEEE, NOIDA, India, 456–462.

[6] Gagan Bansal, Jennifer Wortman Vaughan, Saleema Amershi, Eric Horvitz, Adam Fourney, Hussein Mozannar, Victor Dibia, and Daniel S Weld. 2024. Challenges in human-agent communication. *arXiv preprint arXiv:2412.10380* (2024). https://arxiv.org/abs/2412.10380

[7] Doug A Bowman, Christopher J Rhoton, and Marcio S Pinho. 2002. Text input techniques for immersive virtual environments: An empirical comparison. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 46. SAGE Publications Sage CA: Los Angeles, CA, 2154–2158.

[8] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[9] Virginia Braun and Victoria Clarke. 2023. Toward good practice in thematic analysis: Avoiding common problems and be (com) ing a knowing researcher. *International journal of transgender health* 24, 1 (2023), 1–6.

[10] Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. *How people use chatgpt*. Technical Report. National Bureau of Economic Research.

[11] Ana Paula Chaves and Marco Aurelio Gerosa. 2018. Single or multiple conversational agents? An interactional coherence comparison. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–13.

[12] Andy Cockburn and Bruce McKenzie. 2002. Evaluating the effectiveness of spatial memory in 2D and 3D physical and virtual environments. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, Minneapolis Minnesota USA, 203–210.

[13] Robbe Cools, Matt Gottsacker, Adalberto Simeone, Gerd Bruder, Greg Welch, and Steven Feiner. 2022. Towards a desktop-ar prototyping framework: Prototyping cross-reality between desktops and augmented reality. In *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, IEEE, Singapore, Singapore, 175–182.

[14] Juliet M Corbin and Anselm Strauss. 1990. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology* 13, 1 (1990), 3–21.

[15] Mayukh Das, Phillip Odom, Md. Rakibul Islam, Janardhan Rao (Jana) Doppa, Dan Roth, and Sriraam Natarajan. 2019. Planning with Actively Eliciting Preferences. *Knowledge-Based Systems* 165 (2019), 219–227.

[16] Joe Davison, Joshua Feldman, and Alexander Rush. 2019. Commonsense Knowledge Mining from Pretrained Models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 1173–1178.

[17] I. De Zarzà, J. De Curtò, Gemma Roig, and Carlos T. Calafate. 2023. Optimized Financial Planning: Integrating Individual and Cooperative Budgeting Models with LLM Recommendations. *AI* 5, 1 (2023), 91–114.

[18] Cyan DeVeaux, Eugy Han, Zora Hudson, Jordan Egelman, James A Landay, and Jeremy N Bailenson. 2025. Black immersive virtuality: Racialized experiences of avatar embodiment and customization among black users in social VR. *Computers*

in Human Behavior 168 (2025), 108639.

[19] Neel Dhanaraj, Minseok Jeon, Jeon Ho Kang, Stefanos Nikolaidis, and Satyandra K. Gupta. 2024. Preference Elicitation and Incorporation for Human-Robot Task Scheduling. In *2024 IEEE 20th International Conference on Automation Science and Engineering (CASE)*. IEEE, Bari, Italy, 3103–3110.

[20] Maria Fox and Derek Long. 2002. PDDL+: Modeling continuous time dependent effects. In *Proceedings of the 3rd International NASA Workshop on Planning and Scheduling for Space*, Vol. 4. Citeseer, 34.

[21] Alan D. Fraser, Isabella Branson, Ross C. Hollett, Craig P. Speelman, and Shane L. Rogers. 2024. Do Realistic Avatars Make Virtual Reality Better? Examining Human-like Avatars for VR Social Interactions. *Computers in Human Behavior: Artificial Humans* 2, 2 (2024), 100082.

[22] Ilche Georgievski and Marco Aiello. 2015. HTN Planning: Overview, Comparison, and Beyond. *Artificial Intelligence* 222 (2015), 124–156.

[23] Alfonso Gerevini and Derek Long. 2006. Preferences and soft constraints in PDDL3. In *ICAPS workshop on planning with preferences and soft constraints*. Glasgow, United Kingdom, 46–53.

[24] Alfonso E. Gerevini, Patrik Haslum, Derek Long, Alessandro Saetti, and Yannis Dimopoulos. 2009. Deterministic Planning in the Fifth International Planning Competition: PDDL3 and Experimental Evaluation of the Planners. *Artificial Intelligence* 173, 5 (2009), 619–668.

[25] Moshe Glickman and Tali Sharot. 2025. How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour* 9, 2 (2025), 345–359.

[26] Zhengya Gong and Georgi V. Georgiev. 2020. LITERATURE REVIEW: EXISTING METHODS USING VR TO ENHANCE CREATIVITY. In *Proceedings of the Sixth International Conference on Design Creativity (ICDC 2020)*. The Design Society, 117–124.

[27] Lin Guan, Karthik Valmeekam, Sarath Sreedharan, and Subbarao Kambhampati. 2023. Leveraging pre-trained large language models to construct and utilize world models for model-based task planning. *Advances in Neural Information Processing Systems* 36 (2023), 79081–79094.

[28] Jérôme Guegan, Julien Nelson, and Todd Lubart. 2017. The Relationship Between Contextual Cues in Virtual Environments and Creative Processes. *Cyberpsychology, Behavior, and Social Networking* 20, 3 (2017), 202–206.

[29] Kunal Gupta, Ryo Hajika, Yun Suen Pai, Andreas Duenser, Martin Lochner, and Mark Billinghurst. 2019. In AI We Trust: Investigating the Relationship between Biosignals, Trust and Cognitive Load in VR. In *25th ACM Symposium on Virtual Reality Software and Technology*. ACM, Parramatta NSW Australia, 1–10.

[30] Patsy Healey. 2003. Collaborative Planning in Perspective. *Planning Theory* 2, 2 (2003), 101–123.

[31] Qiaosong Hei, Tianyu Yang, Weihua Dong, Bing He, and Donghui Han. 2025. Leveraging psychology and neuroscience for geospatial cognition research. *Annals of GIS* 31, 1 (2025), 1–13.

[32] Malte Helmert. 2009. Concise Finite-Domain Representations for PDDL Planning Tasks. *Artificial Intelligence* 173, 5-6 (2009), 503–535.

[33] Daniel Höller, Gregor Behnke, Pascal Bercher, Susanne Biundo, Humbert Fiorino, Damien Pellier, and Ron Alford. 2020. HDDL: An extension to PDDL for expressing hierarchical planning problems. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 9883–9891.

[34] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=VtmBAGCN7o

[35] R. Howey, D. Long, and M. Fox. 2004. VAL: Automatic Plan Validation, Continuous Effects and Mixed Initiative Planning Using PDDL. In *16th IEEE International Conference on Tools with Artificial Intelligence*. IEEE Comput. Soc, Boca Raton, FL, USA, 294–301.

[36] Erzhen Hu, Yanhe Chen, Mingyi Li, Vrushank Phadnis, Pingmei Xu, Xun Qian, Alex Olwal, David Kim, Seongkook Heo, and Ruofei Du. 2025. DialogLab: Authoring, Simulating, and Testing Dynamic Human-AI Group Conversations. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. ACM, Busan Republic of Korea, 1–20.

[37] Theodore Jensen, Mohammad Maifi Hasan Khan, Md Abdullah Al Fahim, and Yusuf Albayram. 2021. Trust and Anthropomorphism in Tandem: The Interrelated Nature of Automated Agent Appearance and Reliability in Trustworthiness Perceptions. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference (DIS '21)*. ACM, Virtual Event USA, 1470–1480.

[38] Julia M Juliano, Nicolas Schweighofer, and Sook-Lei Liew. 2022. Increased cognitive load in immersive virtual reality during visuomotor adaptation is associated with decreased long-term retention and context transfer. *Journal of NeuroEngineering and Rehabilitation* 19, 1 (2022), 106.

[39] Farah Juma, Eric I. Hsu, and Sheila A. McIlraith. 2012. Preference-Based Planning via MaxSAT. In *Advances in Artificial Intelligence*. Springer Berlin Heidelberg, Berlin, Heidelberg, 109–120.

[40] Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Paul Saldyt, and Anil B Murthy. 2024. Position: LLMs can't plan, but can help planning in LLM-modulo frameworks. In *Forty-first International Conference on Machine Learning*.

[41] Shyam Sundar Kannan, Vishnunandan L. N. Venkatesh, and Byung-Cheol Min. 2024. SMART-LLM: Smart Multi-Agent Robot Task Planning Using Large Language Models. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Abu Dhabi, United Arab Emirates, 12140–12147.

[42] Callie Y. Kim, Christine P. Lee, and Bilge Mutlu. 2024. Understanding Large-Language Model (LLM)-Powered Human-Robot Interaction. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, Boulder CO USA, 371–380.

[43] Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*. PMLR, PMLR, 5530–5540.

[44] Nils Knoth, Antonia Tolzin, Andreas Janson, and Jan Marco Leimeister. 2024. AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence* 6 (2024), 100225. doi:10.1016/j.caeai.2024.100225

[45] Michal Kolomaznik, Vladimir Petrik, Michal Slama, and Vojtech Jurik. 2024. The role of socio-emotional attributes in enhancing human-AI collaboration. *Frontiers in psychology* 15 (2024), 1369957.

[46] Z. Kootbally, C. Schlenoff, C. Lawler, T. Kramer, and S.K. Gupta. 2015. Towards Robust Assembly with Knowledge Representation for the Planning Domain Definition Language (PDDL). *Robotics and Computer-Integrated Manufacturing* 33 (2015), 42–55.

[47] Huy Viet Le, Sven Mayer, and Niels Henze. 2020. Imprint-Based Input Techniques for Touch-Based Mobile Devices. In *Proceedings of the 19th International Conference on Mobile and Ubiquitous Multimedia* (Essen, Germany) *(MUM '20)*. Association for Computing Machinery, New York, NY, USA, 32–41. doi:10.1145/3428361.3428393

[48] Teven Le Scao and Alexander Rush. 2021. How Many Data Points Is a Prompt Worth?. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 2627–2636.

[49] Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. ACM, New Orleans LA USA, 1–19.

[50] James R Lewis, Brian S Utesch, and Deborah E Maher. 2013. UMUX-LITE: when there's no time for the SUS. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, Paris France, 2099–2102.

[51] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems* 36 (2023), 51991–52008.

[52] Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. 2024. A Survey on LLM-based Multi-Agent Systems: Workflow, Infrastructure, and Challenges. *Vicinagearth* 1, 1 (2024), 9.

[53] Ziming Li, Pinaki Prasanna Babar, and Roshan L Peiris. 2025. Generative Role-Play Communication Training in Virtual Reality for Autistic Individuals: A Study on Job Coach Experiences in Vocational Training Programs. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 591, 22 pages. doi:10.1145/3706598.3713507

[54] Ning Ma, Ruslana Khynevych, Yunqiang Hao, and Yahui Wang. 2025. Effect of anthropomorphism and perceived intelligence in chatbot avatars of visual design on user experience: accounting for perceived empathy and trust. *Frontiers in Computer Science* 7 (2025), 1531976.

[55] Tanja Magoc, Martine Ceberio, and Francois Modave. 2011. Using Preference Constraints to Solve Multi-Criteria Decision Making Problems. *Reliab. Comput.* 15, 3 (2011), 218–229.

[56] Ariel Monteserin and Analía Amandi. 2011. Argumentation–Based Negotiation Planning for Autonomous Agents. *Decision Support Systems* 51, 3 (2011), 532–548.

[57] Catherine S. Oh, Jeremy N. Bailenson, and Gregory F. Welch. 2018. A Systematic Review of Social Presence: Definition, Antecedents, and Implications. *Frontiers in Robotics and AI* 5 (2018), 114.

[58] Vishal Pallagani, Bharath Chandra Muppasani, Kaushik Roy, Francesco Fabiano, Andrea Loreggia, Keerthiram Murugesan, Biplav Srivastava, Francesca Rossi, Lior Horesh, and Amit Sheth. 2024. On the Prospects of Incorporating Large Language Models (LLMs) in Automated Planning and Scheduling (APS). *Proceedings of the International Conference on Automated Planning and Scheduling* 34 (2024), 432–444.

[59] Fumio KISHINO Paul MILGRAM. 1994. A Taxonomy of Mixed Reality Visual Displays. *IEICE TRANSACTIONS on Information* E77-D, 12 (December 1994), 1321–1329.

[60] Gustav Bøg Petersen, Valdemar Stenberdt, Richard E. Mayer, and Guido Makransky. 2023. Collaborative generative learning activities in immersive virtual reality increase learning. *Computers & Education* 207 (2023), 104931. doi:10.1016/j.

compedu.2023.104931

[61] Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, et al. 2023. Chatdev: Communicative agents for software development. *arXiv preprint arXiv:2307.07924* 1 (2023), 15174–15186.

[62] Rivu Radiah, Daniel Roth, Florian Alt, and Yomna Abdelrahman. 2023. The influence of avatar personalization on emotions in vr. *Multimodal Technologies and Interaction* 7, 4 (2023), 38.

[63] Natraj Raman, Sameena Shah, Tucker Balch, and Manuela Veloso. 2021. ViziTex: Interactive visual sense-making of text corpora. In *Proceedings of the second workshop on data science with human in the loop: language advances*. Association for Computational Linguistics, Online, 16–23.

[64] Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*. ACM, Yokohama Japan, 1–7.

[65] Dimitrios Saredakis, Ancret Szpak, Brandon Birckhead, Hannah AD Keage, Albert Rizzo, and Tobias Loetscher. 2020. Factors associated with virtual reality sickness in head-mounted displays: a systematic review and meta-analysis. *Frontiers in human neuroscience* 14 (2020), 96.

[66] Tom Silver, Varun Hariprasad, Reece S Shuttleworth, Nishanth Kumar, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. 2022. PDDL planning with pretrained large language models. In *NeurIPS 2022 foundation models for decision making workshop*.

[67] Mel Slater and Maria V. Sanchez-Vives. 2016. Enhancing Our Lives with Immersive Virtual Reality. *Frontiers in Robotics and AI* Volume 3 - 2016 (2016). doi:10.3389/frobt.2016.00074

[68] David Smith. 2012. Planning as an Iterative Process. *Proceedings of the AAAI Conference on Artificial Intelligence* 26, 1 (2012), 2180–2185.

[69] Shirin Sohrabi and Sheila A McIlraith. 2008. On planning with preferences in HTN. In *Proc. of the 12th Int'l Workshop on Non-Monotonic Reasoning (NMR)*. sn, 241–248.

[70] Chan Hee Song, Brian M. Sadler, Jiaman Wu, Wei-Lun Chao, Clayton Washington, and Yu Su. 2023. LLM-Planner: Few-Shot Grounded Planning for Embodied Agents with Large Language Models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2986–2997.

[71] Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2024. Preference Ranking Optimization for Human Alignment. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 17 (2024), 18990–18998.

[72] Tianqi Song, Yugin Tan, Zicheng Zhu, Yibin Feng, and Yi-Chieh Lee. 2025. Multi-agents are social groups: Investigating social influence of multiple agents in human-agent interactions. *Proceedings of the ACM on Human-Computer Interaction* 9, 7 (2025), 1–33.

[73] Shayan Talaei, Meijin Li, Kanu Grover, James Kent Hippler, Diyi Yang, and Amin Saberi. 2025. StorySage: Conversational Autobiography Writing Powered by a Multi-Agent Framework. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. ACM, Busan Republic of Korea, 1–26.

[74] Su-Mae Tan and Tze Wei Liew. 2022. Multi-chatbot or Single-chatbot? The effects of M-Commerce chatbot interface on source credibility, social presence, trust, and purchase intention. *Human Behavior and Emerging Technologies* 2022, 1 (2022), 2501538.

[75] Liang Tang and Masooda Bashir. 2023. Do We Trust Embodied Agents Who Look Like Us?. In *2023 International Conference on Intelligent Metaverse Technologies & Applications (iMETA)*. IEEE, Tartu, Estonia, 01–07.

[76] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. 2024. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour* 8, 12 (2024), 2293–2303.

[77] Yueyang Wang, Yahui Wang, Xiaoqiong Li, Chengyi Zhao, Ning Ma, and Zixuan Guo. 2023. A comparative study of the typing performance of two mid-air text input methods in virtual environments. *Sensors* 23, 15 (2023), 6988.

[78] Florian Weidner, Gerd Boettcher, Stephanie Arevalo Arboleda, Chenyao Diao, Luljeta Sinani, Christian Kunert, Christoph Gerhardt, Wolfgang Broll, and Alexander Raake. 2023. A Systematic Review on the Visualization of Avatars and Agents in AR & VR Displayed Using Head-Mounted Displays. *IEEE Transactions on Visualization and Computer Graphics* 29, 5 (2023), 2596–2606.

[79] Hao Wen, Yuanchun Li, Guohong Liu, Shanhui Zhao, Tao Yu, Toby Jia-Jun Li, Shiqi Jiang, Yunhao Liu, Yaqin Zhang, and Yunxin Liu. 2024. AutoDroid: LLM-powered Task Automation in Android. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*. ACM, Washington D.C. DC USA, 543–557.

[80] N. Wenk, J. Penalver-Andres, K. A. Buetler, T. Nef, R. M. Müri, and L. Marchal-Crespo. 2023. Effect of Immersive Visualization Technologies on Cognitive Load, Motivation, Usability, and Embodiment. *Virtual Reality* 27, 1 (2023), 307–331.

[81] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang (Eric) Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Ahmed Awadallah, Ryen W. White, Doug Burger, and Chi Wang. 2024. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. In *COLM 2024*. https://www.microsoft.com/en-us/research/publication/autogen-enabling-next-gen-llm-applications-via-multi-agent-conversation-framework/

[82] Tongshuang Wu, Ellen Jiang, Aaron Donsbach, Jeff Gray, Alejandra Molina, Michael Terry, and Carrie J Cai. 2022. Promptchainer: Chaining large language model prompts through visual programming. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. ACM, New Orleans LA USA, 1–10.

[83] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. ACM, New Orleans LA USA, 1–22.

[84] Yuanyuan Xu, Weiting Gao, Yining Wang, Xinyang Shan, and Yin-Shan Lin. 2024. Enhancing User Experience and Trust in Advanced LLM-based Conversational Agents. *Computing and Artificial Intelligence* 2, 2 (2024), 1467.

[85] Xinyi Yan, Chengxi Luo, Charles L. A. Clarke, Nick Craswell, Ellen M. Voorhees, and Pablo Castells. 2022. Human Preferences as Dueling Bandits. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Madrid Spain, 567–577.

[86] Saelyne Yang, Jo Vermeulen, George Fitzmaurice, and Justin Matejka. 2024. AQuA: Automated Question-Answering in Software Tutorial Videos with Visual Anchors. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 928, 19 pages. doi:10.1145/3613904.3642752

[87] Kai Zhang, Bernal Jimenez Gutierrez, and Yu Su. 2023. Aligning Instruction Tasks Unlocks Large Language Models as Zero-Shot Relation Extractors. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, Toronto, Canada, 794–812.

[88] Nan Zhang, Hongkai Wang, Tianqi Huang, Xinran Zhang, and Hongen Liao. 2024. A VR Environment for Human Anatomical Variation Education: Modeling, Visualization and Interaction. *IEEE Transactions on Learning Technologies* 17 (2024), 391–403.

[89] Shuning Zhang, Hui Wang, and Xin Yi. 2025. Exploring Collaboration Patterns and Strategies in Human-AI Co-creation through the Lens of Agency: A Scoping Review of the Top-tier HCI Literature. *Proceedings of the ACM on Human-Computer Interaction* 9, 7 (2025), 1–43.

[90] Zirui Zhao, Wee Sun Lee, and David Hsu. 2023. Large language models as commonsense knowledge for large-scale task planning. *Advances in Neural Information Processing Systems* 36 (2023), 31967–31987.

[91] Zhehua Zhou, Jiayang Song, Kunpeng Yao, Zhan Shu, and Lei Ma. 2024. ISR-LLM: Iterative Self-Refined Large Language Model for Long-Horizon Sequential Task Planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, Yokohama, Japan, 2081–2088.