Research Proposal

Towards Collaborative Spatial Intelligence: Multi-Agent Systems for User-Guided VR World Building

Motivation

Creating custom VR scenes inherits the idea of world building and brings it to 3D immersive space from traditional 2D environments (like Minecraft). This evolution significantly expands the boundaries of Reality Technology and opens new possibilities for human creativity and expression.

Limitations of existing systems

Limitations of current spatial understanding systems include:

1. Relatively poor spatial reference handling: current systems struggle in basic user-related commands like “place the table next to me”, often misinterpreting directional references and user-relative positioning.
2. Lack of real-time spatial adaptation: when users make adjustments in the VR environments, existing systems cannot dynamically adjust object placement relative to the user's new position, breaking immersion and spatial coherence.
3. Limited Multi-scale spatial reasoning: existing systems cannot handle both scene & room scale spatial planning and object-scale positioning in a seemingless workflow.

Experiment propose

I propose a novel Multi-Agent System for intelligent 3D asset placement that transforms this paradigm. Through natural language interaction, users can express their creative vision in simple prompts (e.g., "Make the coffee shop layout more spatially for friends to meet", “place the table next to me”), while specialized AI agents collaborate to understand the **ambiguous user request**, intelligently select appropriate assets, and autonomously place them in spatially coherent arrangements according to ML predictions.

Importance of spatial understanding system

Why is spatial understanding important?

1. Spatial understanding systems with user view & language input can power up robots to do arrangements in the physical world. For example, with multiple housekeeping robots, one prompt from the user (eg., “make the layout of the current room looks more spatial”) will let robots clean the room on their own.
2. Also, specific prompts like (“place the table here”, “give the coffee cup to me”, etc) will enable robots to do tasks for people with disabilities, making their lives easier.
3. A spatial understanding system can be a better option for AI to collaborate with people in fulfilling real-world tasks like rescues and surgeries.

Literature Review

1. Article: How People Prompt Generative AI to Create Interactive VR Scenes

Research on "How People Prompt Generative AI to Create Interactive VR Scenes" identified four key user expectations: (1) embodied knowledge of environment, (2) understanding of embodied prompts, (3) recall of previous states, and (4) commonsense understanding of objects ArXivAcm

Key Finding: Users prompt differently when in situ (within VR) vs ex situ (outside VR)
System: Ostaad - single conversational programming agent for VR scene creation

Limitation:

1. Execution of prompts which include spatial references like “put this to the right of that” is difficult, which means it lacks precision. (maybe require multi-model approach)
2. Cannot understand vague commands like “put the chair over there” compared to normal humans. (need use user’s view as a pointer for reference)

Link: https://dl.acm.org/doi/pdf/10.1145/3643834.3661547

2. Article: Social Conjuring: Multi-User Runtime Collaboration with AI in Building Virtual 3D Worlds

This article focuses on **Multi-user** collaborative creation of interactive VR scenes for social purposes. One of the research questions worth further investigation is “how to develop a real-time, collaborative, spatially-aware system, integrating LLM and VLM for co-creation of the VR world?” The proposed social conjuring system utilizes multi-model to accomplish this. One special mechanism is the “Decider” module which takes user

prompts as input, and output becomes binary (whether categorized as “static” or “interactive” prompt)

Limitation:

1. The system integrates a **Spatial Reasoning submodule** that decides how objects should be **placed, oriented, and scaled** in a scene. However, since it is a one-time generation tool, it cannot focus on precise scene adjustment (eg., “put the cup on my hand”). Instead, it can only handle simple modifications like (“add a unicorn in the zoo created).
2. It focuses more on the scene generation and **multi-user collaboration** instead of precise spatial layout within the scene and precise scene modification, which makes it more suitable for elaborating creativity instead of real-world usage.

Link: https://arxiv.org/html/2410.00274v2

Potential Research Questions

1. Can multi-agent collaboration produce better spatial layouts than single-agent systems?
2. How do specialized spatial agents (layout, relationships, validation) work together?
3. Can agents understand vague prompts from natural language? (eg., “place the table here", “make the room layout cleaner and more spatial”)

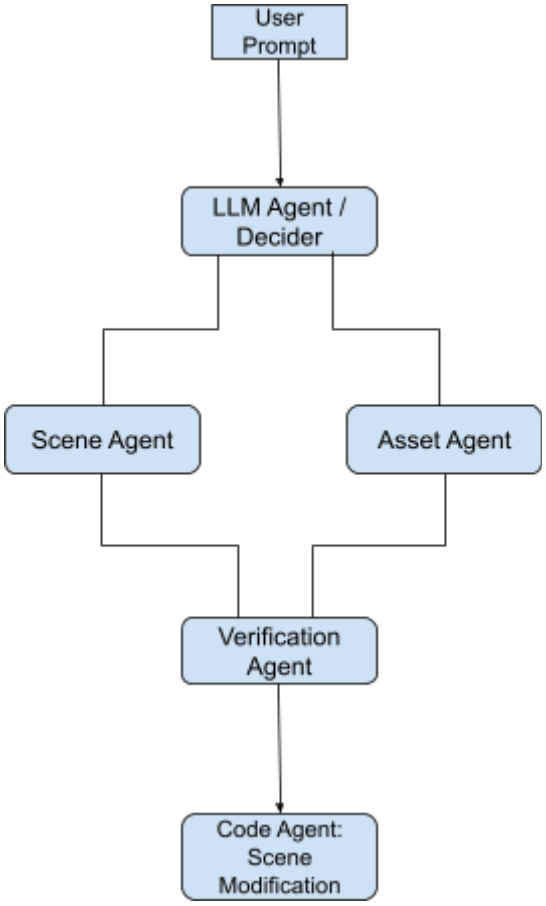
PS: The user prompt should be as simple as possible, like: “create a cozy coffee shop, add a sofa beside me, etc.”.

Pipeline

Basically there are 5 agents:

1. The language agent to process user prompt (LLM + decider);
2. The scene agent to handle scene-level determination of the object (eg., the layout of the space). The scene agent helps determine the existing object within the scene that is mentioned in the user prompt. (move the table beside the window);
3. The asset agent to handle object-level generation, modification in smaller scope (example prompt: create a sofa beside me);
4. Verification agent checks whether the decision made from scene and asset agents makes sense or not;
5. Code agent generates code to make the actual change.

Workflow:



Spatial Reasoning & Inference

This is the most important part of the project and it sets our research aside from the previous ones - It is about collaborative spatial intelligence among agents.

This means that each agent focuses **part** of the spatial intelligence task assigned by the user-prompt. For example, one agent (scene agent) determines the basic layout of the room, another agent (asset agent) determines the exact placement of each asset within the space. A verification agent checks whether the result fulfills the expectation. If not, then the scene agent and asset agent should work again on their task. After several iterations, these agents pass the decision to the code agent to generate code to actually change the position of each asset of the space. As a result, all agents collaborate in the backend to determine the spatial relationship between user and scene, user and assets, scene and assets in fulfilling the task.

1. Compared to most existing content creation tools that rely on 2D screens, this directly lets users see the result of their creative idea on a real-world scale with just language prompts, which lowers the difficulty of content creation and makes the result more intuitive and straightforward to eye view.

2. Compared to previous researches in VR content generation, (Multi-Agent Spatial Collaboration vs. Single-Agent Systems) current systems like Ostaad suffer from poor spatial reference handling ("spatial references to objects 'to the right of' or 'in front of' did not generally work well"). Our distributed approach addresses this limitation through specialized agent coordination, enabling real-time spatial understanding with user-relative positioning.
   Relative article: *How People Prompt Generative AI to Create Interactive VR Scenes*
   Link: https://dl.acm.org/doi/fullHtml/10.1145/3643834.3661547

3. Compared to previous research in spatial reasoning work, existing spatial reasoning research focuses on object to object relationships. Current 3D-based LLMs "still fall short in situated understanding, a fundamental capability for completing embodied tasks". This multi-agent spatial reasoning system uniquely maintains user position awareness, enabling natural commands like "place the chair beside me" through ego-centric spatial modeling. (need the agent to gather spatial information from headset camera to do the calculation and prediction)
   Relative article: *Situational Awareness Matters in 3D Vision Language Reasoning" - States that current 3D-based LLMs*
   Link: https://arxiv.org/pdf/2406.0754

User Study

Compare it with traditional & existing 3D scene generation models to test:

1. Spatial understanding precision.
2. User feedback in feasibility and usability (needs further investigation and design to maintain the effectiveness and the research scope)

Novelty

There are some novel aspects of this research topic which worth to discuss:

1. Unlike traditional single-agent systems to handle 3D scene generation tasks, this utilizes specialized multi-agent systems where each agent contributes distinct spatial reasoning expertise. Also good for multi-GPU deployment (scalability boost).
2. First system to understand and maintain spatial relationships between embodied user position and virtual objects from vague prompts according to AI ("beside me," "in front of me") through agent collaboration, enabling dynamic spatial placement based on user location in VR space.

Limitations

1. Limited number of agents, cannot handle complex spatial tasks, especially animation tasks like (“make this table jump to me”).
2. Requires large scale GPU deployment. May go-over this my having scene by default, so the user input a scene at first (eg., coffee shop, bedroom), and then make the creation and modification of the asset within the scene in real-time.
3. Cannot handle prompts that are “too” vague as human (like “leave the table here”). May need to use the user’s eye as a pointer for reference for AI to determine the location which “here” refers to.

Existing Questions to Consider

1. In order to focus on the spatial understanding, will it be better to have a pre-defined scene? For example, a bedroom scene by default will help eliminate the time for generating an actual scene in real-time. We can prepare several scenes by default (like coffee shop, bedroom, supermarket). (check)

2. Regarding spatial understanding, it might be more innovative to design a MAS (Multi-Agent System) with **collaborative spatial understanding**. That means, with each agent specifying & professional in part of the spatial understanding task. Instead of the linear return passing of the current workflow, agents should be able to dynamically discuss with each other to improve the result. (focus on the collaborative spatial understanding) (check)

Paper Submission & Timeline

1. Targeting the IEEE VR deadline Mid-September 2026 for the full-paper submission
2. First submit the 2-page DEMO to UIST deadline Late-June to Mid-July
3. 1-hour weekly meeting for the research group.

What should I do next?

What to do next for me:

1. Refine the existing proposal to make it more clear to the focus. (priority!!!)
2. Investigate the multi-agent system (the system design part)
    - How does the agent integrate to change the code of the project?
    - Refining the scope for what kind of code need to be changed to avoid total crush (for example, if the agent decide to delete code “import from three.js…”, then the whole scene black out, we should definitely avoid this)
3. Investigate the spatial understanding theory (for prompt design)
    - LSTM, SLAM model (these are traditional 3D CV ML models, although may not be relevant for Multi-Agent System)
    - Especially focus on the LLM driven spatial model, or LLM with spatial module integrated.