CARNEGIE MELLON UNIVERSITY

School of Architecture

College of Fine Arts

Thesis

Submitted in Partial Fulfillment of the requirements for the degree of

Master of Science in Computational Design

TITLE:

Exploring the Convergence of AI and VR in Architectural Design, From Ideation to Actionable Outputs

AUTHOR:

Zhaowei Chen

ACCEPTED BY ADVISORY COMMITTEE:

Joshua D. Bard		May 6th, 2025
Professor NAME	Principal Advisor	DATE
Jimmy Wei-Chun Cheng		May 6th, 2025
Professor NAME	Advisor	DATE

Exploring the Convergence of AI and VR in Architectural Design, From Ideation to Actionable Outputs

Zhaowei Chen
Advisors: Joshua D. Bard, Jimmy Wei-Chun Cheng

May 1, 2025

Abstract

The integration of Artificial Intelligence (AI) and Virtual Reality (VR) presents new opportunities for transforming architectural design workflows, particularly in small-scale projects. This research introduces MerGen, a prototype system that combines generative AI models, large language models (LLMs), and immersive VR to support real-time collaboration between designers and intelligent agents. By situating users within panoramic environments, MerGen enables continuous spatial awareness and interactive control throughout the design process.

Generative models like Stable Diffusion allow for real-time visual modifications based on user intent, while LLMs provide contextual suggestions, semantic dialogue, and design documentation. This workflow bridges conceptual ideation with actionable outputs, including material recommendations and spec-style summaries.

A modular, experiment-driven methodology is used to evaluate MerGen's effectiveness across two key stages of design: early sketching and thematic exploration, and material selection and specification. The study examines how AI–VR integration enhances ideation, visualization, and decision-making. Findings highlight both the promise and limitations of this approach, offering insights into its role in supporting immersive, multimodal architectural workflows.



Contents

1 Introduction **4**

1.1 AI in Architectural Design 4

1.1.1 The Rise of Artificial Intelligence in Architectural Design 4

1.1.2 AI Adoption and Concerns in the Architectural Profession 4

1.2 The Evolution of Virtual Reality in Architectural Design 4

1.3 The Convergence of AI and VR in Architectural Design 5

1.4 Research Question 6

1.5 Structure of the Thesis 6

2 Background **8**

2.1 Generative AI: Potential and Limitations 8

2.2 VR in Architectural Design: Spatial Awareness but Limited Interactivity 9

2.3 AI–VR Integration Research 10

2.4 Positioning of This Study 11

3 Hypothesis **12**

4 Methods **13**

4.1 Self-Exploration of Generative AI for Panoramic Image Generation 13

4.1.1 Experiment with Skybox AI for Panoramic Image Generation 13

4.1.2 Initial Test with Stable Diffusion for Panoramic Image Generation 14

4.1.3 Experiment with Stable Diffusion Inpainting 16

4.1.4 AI-Assisted Design Advisory Using LLaMA 3.2 Vision Model 17

4.1.5 Summary of Findings from Self-Exploration 19

4.2 Technical Details of System Development 20

4.2.1 Unity-Based VR Frontend 22

4.2.2 Backend AI Framework 27

4.2.3 31

4.3 Early Sketching and Thematic Exploration 32

4.3.1 Research Goals and Rationale 32

4.3.2 Experimental Workflow 33

4.3.3 Detailed Experimental Narratives 35

4.3.4 Key Findings from Experiments 46

4.3.5 Addressing AIA survey Concerns 48

4.4 Material Selection and Specification 48

4.4.1 Research Goals and Rationale 48

4.4.2 Methodology and Workflow 49

4.4.3 Detailed Experimental Narratives 50

4.4.4 Key Findings from Experiments 57

4.4.5 Addressing AIA survey Concerns 58

5 Results **60**

5.1 Preliminary Findings 60

5.2 Limitations 61

6 Discussion and Conclusion **63**

6.1 Contributions 63

6.2 Next Steps 63

1 Introduction

1.1 AI in Architectural Design

1.1.1 The Rise of Artificial Intelligence in Architectural Design

Artificial Intelligence (AI) has rapidly emerged as a transformative force in architectural design, driven by advancements in machine learning, generative AI, and automation tools. In the Architecture, Engineering, and Construction (AEC) industry, AI technologies are increasingly being leveraged to optimize design workflows, enhance creative exploration, and streamline project execution[27].

One of the most impactful applications of AI in architecture is Generative AI, which enables the automatic generation of design alternatives based on predefined parameters. Techniques such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Diffusion Models are widely used to generate architectural forms, optimize spatial layouts, and explore aesthetic variations [23]. AI-driven tools, such as Stable Diffusion, MidJourney, and DALL-E, allow architects to quickly visualize complex design ideas, reducing the time needed for conceptual development [5].

Furthermore, AI has revolutionized the architectural visualization process, improving rendering efficiency and real-time simulations. AI-driven rendering engines can generate high-quality visualizations without requiring extensive computational resources, making design iterations more accessible and cost-effective [34]. Architectural firms, including Zaha Hadid Architects (ZHA) and MVRDV, have begun integrating AI-generated concepts into their workflows to explore new design languages and material possibilities [23].

1.1.2 AI Adoption and Concerns in the Architectural Profession

A recent survey by the American Institute of Architects (AIA) highlights a growing interest in AI in the architectural profession, but significant concerns remain about its implementation [26]. The report reveals that while more than half (53%) of architects have experimented with AI, only 6% use it regularly in their practice. Furthermore, 8% of companies have integrated AI into their workflow, and larger companies are leading adoption efforts [26].

Despite this interest, architects express a wide range of concerns about AI. The most frequently cited issues include inaccuracies in AI-generated output (64%) and unintended consequences (64%). This cautious approach has slowed the adoption of AI, as many firms remain hesitant due to potential challenges.

Given these findings, it is evident that for AI adoption to grow in the architectural industry, these concerns must be systematically addressed in the development of AI-driven applications. Future AI tools should prioritize accuracy, transparency, and ethical considerations to align with the needs of the profession. By mitigating these risks, the industry can move toward broader and more responsible AI integration.

1.2 The Evolution of Virtual Reality in Architectural Design

Virtual Reality (VR) has emerged as a powerful tool for architectural visualization, immersive design exploration, and collaborative project development[33]. Unlike traditional Computer-Aided Design (CAD)

tools, VR enables designers to experience architectural spaces at a 1:1 scale, enhancing their ability to assess spatial relationships, materiality, and user experience[21].

One of the key advantages of VR is its ability to improve real-time collaboration. Architects, clients, and stakeholders can participate in immersive walk-throughs, allowing them to make informed decisions before construction begins[36]. Studies indicate that VR-based design reviews can reduce design errors by up to 90%, leading to significant cost savings and improved project efficiency[14]. VR has also been integrated into the Building Information Modeling (BIM) ecosystem, allowing seamless interaction between digital models and immersive environments[33]. Advanced VR systems support real-time rendering and physics-based simulations, allowing architects to analyze lighting conditions, acoustics, and human movement within virtual spaces[21].

1.3 The Convergence of AI and VR in Architectural Design

The integration of AI and VR presents a paradigm shift in architectural workflows, bridging the gap between computational design and immersive experience [14]. While AI provides automation, predictive analytics, and generative capabilities, VR acts as an interactive medium, allowing architects to engage with AI-generated content in a spatially intuitive manner. This synergy can mitigate some of the key concerns architects have about AI, particularly inaccuracy, unintended consequences, and usability constraints [26].

One of the primary concerns about AI in architecture is the accuracy of AI-generated outputs [26]. Architects fear that AI-driven designs may produce misleading or unrealistic results. VR can serve as a validation tool, enabling real-time evaluation of AI-generated designs within immersive environments. By placing architects inside full-scale, navigable representations of AI-generated models, VR allows them to detect errors, assess proportions, and validate material selections before implementation [33]. This direct interaction with AI-generated outputs reduces the risk of inaccuracies and enhances trust in AI-driven processes.

Another major issue is the unintended consequences of AI-driven decision-making [26]. AI models, particularly generative ones, may overlook contextual constraints or produce designs that conflict with functional requirements. VR enables designers to conduct real-time spatial testing—for example, by simulating lighting conditions, material behavior, and human interactions within AI-generated spaces [34]. This ensures that AI outputs are not only creative but also functionally viable.

Furthermore, usability concerns have been a barrier to AI adoption, as architects often lack intuitive ways to interact with AI-driven tools [26]. VR introduces natural user interfaces, allowing architects to engage with AI models through gestures, voice commands, and spatial navigation rather than complex coding or parameter-based interactions [14]. This makes AI tools more accessible, particularly for architects without programming expertise.

Some emerging applications of AI-VR integration in architecture include:

Immersive Design Exploration Driven by AI – AI-generated design alternatives can be interactively visualized and modified within VR, allowing designers to explore variations dynamically [27]. This interactive approach enhances design flexibility and reduces the risk of AI-generated errors going unnoticed.

Automated Material and Lighting Simulations – AI algorithms can predict material behavior and simulate real-world lighting conditions within a VR environment, improving design precision [33]. This ensures that AI-generated designs align with real-world physics, addressing concerns about unrealistic AI outputs.

Intelligent Virtual Assistants – AI-powered chatbots and assistants embedded in VR systems can provide design recommendations, real-time feedback, and project documentation support [34]. These assistants can help explain AI-generated suggestions, increasing transparency and reducing concerns about black-box decision-making in AI [26].

The potential of AI-VR collaboration extends beyond visualization, affecting sketching, material selection, and client engagement [14]. However, challenges such as data integration, software interoperability, and computational constraints remain key areas for future research [33]. Addressing these challenges will be crucial in maximizing the benefits of AI-VR integration while ensuring reliability, transparency, and usability in architectural workflows.

1.4 Research Question

Main Question

"How could the convergence of AI with VR impact the architectural design pipeline, from ideation to actionable outputs, in small-scale projects?"

Sub-questions

1. How can generative AI influence real-time interactivity in architectural design?
2. What are the technical and workflow challenges in integrating AI with VR technologies?
3. How can AI-VR platforms produce actionable deliverables like project specification books?

1.5 Structure of the Thesis

This thesis is structured into six main chapters, following a logical progression from background and theoretical exploration to methodology, results, discussion, and conclusion.

Chapter 1 introduces the research topic, providing background on the role of Artificial Intelligence (AI) and Virtual Reality (VR) in architectural design. It highlights the growing interest in AI within the architectural profession while acknowledging the concerns that have hindered its adoption. The chapter then explores how AI-VR integration can address these challenges, presenting the research questions and objectives that guide the study. A brief summary of the methodology and significance of the research is also included.

Chapter 2 provides a literature review of existing research on AI and VR in architecture, discussing their individual applications and limitations. It examines the strengths and weaknesses of generative AI models in design workflows and explores how VR enhances spatial awareness but lacks real-time interactivity. The chapter also positions this research within the broader AI-VR integration landscape, identifying gaps that justify the study.

Chapter 3 formulates the central hypothesis of the study: that AI and VR integration can enhance small-scale architectural workflows by enabling real-time, AI-assisted design interactions. The hypothesis is structured around key areas such as early sketching, thematic exploration, and material selection, outlining how AI-VR collaboration could improve architectural design processes.

Chapter 4 outlines the experimental methodology used in the research. It describes the self-exploration phase, where various AI tools were tested for panoramic image generation, followed by the experimental modules focusing on AI-assisted early sketching, thematic exploration, and material selection in a VR environment. The chapter also details the technical development of an AI-powered backend and a VR-based frontend to facilitate real-time interactions. Additionally, it presents the evaluation metrics used to assess the effectiveness of AI-VR integration in enhancing architectural workflows.

Chapter 5 presents the findings from the experimental modules, analyzing how AI-VR integration influenced design iteration, spatial adaptability, and the generation of actionable outputs. It also discusses technical limitations, such as real-time adaptability challenges and VR-AI data integration issues, which need to be addressed for broader adoption of AI-driven immersive design environments.

Chapter 6 concludes the thesis by reflecting on the contributions of the study, emphasizing how AI-VR integration can transform architectural workflows. It acknowledges limitations, such as scalability issues and the adaptability of AI tools for broader use cases, while also offering recommendations for future research. The chapter suggests improvements in AI-driven immersive design tools, expanded real-world validation, and further exploration of AI-VR integration in different architectural contexts.

This structure ensures a comprehensive exploration of AI and VR in architectural design, guiding the reader through the study's theoretical foundations, experimental process, and findings, ultimately contributing to the discourse on AI-driven innovation in architecture.

2 Background

2.1 Generative AI: Potential and Limitations

Generative Artificial Intelligence (AI) is increasingly integral to the early stages of architectural design. Its ability to transform textual or visual prompts into conceptual imagery allows architects to accelerate ideation and explore diverse design directions. Tools such as Stable Diffusion and DALL·E have demonstrated their capacity to augment creative workflows and expand visual languages in architecture [28]. For instance, Cao et al. [6] conducted a large-scale academic study with over 200 architecture students, illustrating how Stable Diffusion can boost engagement and ideation fluency within studio-based learning environments.

The potential of generative AI lies in its ability to generate stylistically and formally novel outcomes by learning from extensive architectural datasets. This enables even small practices to access high-level creative variation that would traditionally require significant time and resources [3]. However, this strength also exposes a critical weakness: the opaque, "black box" decision-making processes of AI systems limit user understanding and control. In professional settings, this poses ethical and operational challenges—especially in project phases that demand traceability, such as planning submissions or client justification [24]. Gado and Matter's study [24], which analyzed real-world BIM implementations, confirms that successful digital adoption hinges on aligning AI tools with transparent, structured workflows.

Another major constraint is the linearity of most AI design tools. They often operate in a single-step, prompt-output model that does not support real-time, iterative refinement—an essential feature of architectural design processes [32]. Davis et al. [32] explored these limitations in immersive environments, revealing how static AI outputs restrict meaningful interaction in dynamic spatial scenarios. Even in advanced architectural visualization, where AI tools such as SnapRender and Midjourney enhance rendering aesthetics, the emphasis remains on surface realism rather than process integration or design validation [25].

Efforts to overcome these limitations emphasize multimodal interaction and design reasoning. Cheung and Dall'Asta [8] propose a collaborative AI design framework that integrates sketches, voice, and visual inputs to support iterative idea development. Their system lowers technical barriers and enables intuitive dialogue with AI. Zhang et al. [39] further validate this approach through a sketch-based experimental setup, showing how students used AI to extend and reframe initial design ideas, though still struggling with architectural semantics and constructability.

Nonetheless, reliance on AI systems must be managed carefully. Empirical studies show that while AI can catalyze idea generation, overdependence may lead to reduced critical engagement and superficial outcomes. Zeytin et al. [38] documented three distinct user behaviors in a design studio environment: full reliance on AI tools, selective integration, and cautious skepticism. Their findings highlight that students who critically reflected on AI outputs produced more contextually grounded and innovative designs, underscoring the importance of pedagogical strategies that reinforce human agency.

From a human-computer interaction perspective, scholars advocate rethinking AI not as a tool but as a co-creator—capable of adapting to the designer's intent and providing meaningful, explainable support throughout the design process. Cheung and Dall'Asta [9] argue that this requires systems to be transparent, adaptive, and aligned with designers' cognitive workflows. Complementing this, Shi et al. [31] conducted a

meta-review of 93 studies and concluded that successful design–AI collaboration depends on flexibility, user control, and ethical framing—especially in high-stakes, context-sensitive domains like architecture.

In sum, while generative AI introduces new capacities for creativity, speed, and visualization, its integration into architectural workflows must be approached critically. Effective adoption depends not only on technological sophistication but on ethical, pedagogical, and methodological considerations that maintain the centrality of the human designer.

2.2 VR in Architectural Design: Spatial Awareness but Limited Interactivity

Virtual Reality (VR) has become an increasingly valuable tool in architectural design by enhancing spatial perception through immersive visualization. Unlike traditional media, VR enables designers and stakeholders to inhabit architectural spaces at full scale, supporting a richer understanding of spatial proportions, lighting, materiality, and circulation. Davis et al. [32] emphasize that VR allows for experiential engagement with form and movement that 2D drawings cannot provide. Abu Alatta and Freewan [2] further affirm that VR significantly improves students' spatial cognition and their ability to evaluate user experience.

However, in most architectural practices, VR remains primarily a presentation medium. Yu et al. [37] identify that VR tools are commonly confined to the end of the design pipeline and are used to review pre-modelled environments without supporting real-time modifications. Ergün et al. [17] highlight that while VR improves environmental realism and feedback quality, its interactive design support remains limited in conventional workflows.

To counteract these limitations, recent innovations have aimed to integrate VR into conceptual and generative stages. Castelo-Branco and Leitão [7] introduced live algorithmic coding within VR, enabling designers to manipulate geometry in real time. Doma and Şener [12] explored dynamic user-scaling mechanisms, allowing users to ideate spatial compositions directly through embodied interaction. Their later project, Dreamscape Bricks VR [11], confirmed that immersive and hybrid modes support design exploration better than traditional desktop tools.

Participatory and collaborative potentials of VR have also advanced. Lee et al. [22] developed a panoramic AVR interface that supports remote, multi-user evaluation. Hsu et al. [18] proposed a VR studio system enabling object manipulation, role-based interaction, and rollback functions, facilitating real-time critique and iteration during design discussions.

In architectural education, VR has shown strong pedagogical value. Aydin and Aktaş [4] studied two curriculum-wide digital design ecosystems and found that VR enhanced student creativity and design engagement. Agirachman et al. [1] implemented an affordance-based VR system where students could match design features to user needs through immersive feedback. Ibrahim and El Shakhs [20] demonstrated that using VR to evaluate emotional concepts like "levitation" led students to refine spatial qualities such as lighting, materiality, and circulation, aligning design form with experiential goals.

Cho and Suh [10] found that students with lower spatial ability performed better in immersive tests than in static ones, suggesting VR's potential to equalize cognitive disparities. Their results highlight the need for VR-inclusive pedagogical models that support a diversity of learning styles.

In professional contexts, collaborative and interdisciplinary VR platforms are emerging. Roupé et al. [29] developed ViCoDE, a system combining multi-touch tables and HMDs in real-world hospital design. It enabled synchronized interaction across formats and empowered users—from architects to clinicians—to collaboratively test and revise spatial configurations. The system allowed for rapid feedback loops and facilitated spatial negotiations in ways traditional CAD systems could not.

Still, these systems remain the exception rather than the norm. Davis et al. [32] argue that to fully capitalize on VR's design potential, platforms must evolve to support semantic reasoning, dynamic interaction, and continuous feedback—qualities that are only beginning to appear in experimental systems.

In conclusion, while VR enhances spatial comprehension and supports user engagement, its transition from a presentation tool to a co-creative design instrument depends on integration with intelligent systems and interactive workflows. Future research and development should focus on embedding design reasoning and real-time input into immersive environments to transform how architects explore, test, and communicate spatial ideas.

2.3 AI–VR Integration Research

The convergence of Artificial Intelligence (AI) and Virtual Reality (VR) technologies has recently emerged as a pivotal direction in architectural design research. These two domains—AI as a computational engine of design generation and reasoning, and VR as an immersive interface for spatial interaction—hold significant potential when unified into a seamless co-creative environment. While early investigations into AI–VR integration focused on discrete applications such as AI-driven scene composition, speech recognition interfaces, or procedurally generated environments [32], most of these remain conceptually experimental or limited to non-architectural domains like gaming.

In the architectural context, only a limited number of studies have explored truly integrated workflows that support natural language dialogue, contextual spatial feedback, semantic reasoning, and iterative refinement. The few systems that do attempt this are often built as research prototypes rather than deployable design platforms. One of the most advanced among them is the immersive collaborative discussion system by Hsu et al. [18], which incorporates mid-air sketching, gesture-based object manipulation, and a "rollback" function allowing real-time design iteration.

Affordance-based systems such as the one proposed by Agirachman et al. [1] integrate cognitive design support into VR, matching built elements with user-centered design criteria. Aydin and Aktaş [4] explore curriculum-scale ecosystems that could evolve into AI-supported immersive design studios. The Sculptor system by Engeli and Kurmann [16] introduces intelligent agents and reactive components into the design environment, enabling collision detection, spatial navigation assistance, and autonomous layout suggestions—all driven by AI.

Hussain et al. [19] extend this integration into the realm of construction safety, combining ChatGPT with VR to train multilingual migrant workers. The iSafeCom platform enables voice-based interaction, context-aware AI responses, and real-time safety feedback, showing the applicability of AI–VR fusion in real-world, domain-specific environments.

Similarly, Shehadeh and Alshboul [30] unite machine learning, VR, and BIM for clash detection in additive manufacturing. Their system incorporates explainable AI via SHAP analysis, allowing for design validation in immersive environments—marking a practical evolution toward AI-mediated spatial intelligence.

Drogemuller et al. [13] introduce the theoretical and technological construct of Paramersive Design—a fusion of parametric modeling, immersive feedback, and AI reasoning. Their prototypes, Paramodal and Paramaxes, allow voice and gesture-based parametric control and visual-tactile manipulation of design variables, forming an embodied interface for iterative exploration.

Emory and Doss [15] detail a multi-year pedagogical integration of AI-generated content and VR storytelling. Through character-based narrative design, students iteratively shaped immersive environments guided by prompts from ChatGPT and images from Midjourney, demonstrating a compelling case of creative co-authorship between human and AI agents in architectural settings.

Most recently, Wu and Han [35] present a thoroughly evaluated AI–VR interior design platform featuring real-time prop transformation, roaming interaction, and multi-sensory user feedback. Their system architecture integrates a loading interface, virtual tour, and VR interaction modules, allowing users to modify furniture, flooring, ceilings, and artwork using intuitive gaze and handheld gestures. Empirical evaluations showed over 99% interaction sensitivity, sub-second delay times, and stable performance for up to 60 simultaneous users. The use of Fuzzy Comprehensive Evaluation (FCE) methodology further demonstrates the system's robustness and academic rigor, underscoring the readiness of such AI–VR systems for mainstream design practice.

Despite such promising developments, Davis et al. [32] caution that most AI–VR systems still lack robust semantic reasoning, adaptability to spatial logic, and regulatory compliance capabilities—limitations that must be addressed to embed AI meaningfully in professional practice. As architectural design increasingly embraces hybrid digital workflows, future AI–VR systems must be capable of interpreting intent, reacting to context, and learning from spatial behavior to become effective co-creators in architectural processes.

2.4 Positioning of This Study

This thesis investigates the integration of generative Artificial Intelligence (AI)—including large language models (LLMs), image generation tools, and Retrieval-Augmented Generation (RAG) architectures—with immersive Virtual Reality (VR) to create a real-time, multimodal design environment. Within this system, architects engage through voice, gesture, and panoramic vision, enabling not only spatial ideation but also material refinement and specification within the same immersive workflow.

The need for such integration is underscored by the current bifurcation of design processes: AI technologies have proven highly effective for concept generation and stylistic diversity. For example, Ploennigs and Berger [28] illustrate how generative models can transform abstract prompts into spatial imagery, accelerating visual ideation. Amer [3] complements this by showing how architectural language models contribute to hybridized formal development and typological evolution.

Conversely, VR environments have predominantly served as post-design visualization platforms, enhancing spatial comprehension but rarely supporting design iteration. Davis et al. [32] argue that most current systems fail to embed semantic intelligence, limiting user agency and contextual responsiveness. Abu Alatta

and Freewan [2] demonstrate that although VR improves student engagement with spatial experience, its utility is undercut by the absence of real-time, intelligent feedback loops.

This thesis aims to overcome such limitations by embedding AI reasoning into VR interfaces, thus transforming VR from a passive visualization tool into an active, collaborative design partner. The framework draws upon precedents like the Paramodal prototype by Drogemuller et al. [13], which enables voice and gesture-based parametric control, and the Sculptor system by Engeli and Kurmann [16], which introduced autonomous agents for real-time design support and environmental simulation.

To validate practical feasibility, this study follows system design logic tested by Wu and Han [35], whose AI–VR interactive decoration system demonstrated real-time responsiveness, high interaction precision, and user scalability. It also draws inspiration from Hsu et al. [18], who emphasized real-time collaboration through object rollback and sketch manipulation in multi-user design VR studios.

The inclusion of natural language dialogue as a central interface aligns with proven methods in the iSafeCom system, where Hussain et al. [19] combined ChatGPT with immersive scenarios for safety training, resulting in statistically significant improvements in performance and user comprehension. Likewise, Emory and Doss [15] showcase how LLMs like ChatGPT, when integrated into narrative design studios, facilitate co-creative authorship between humans and AI.

This thesis is also methodologically positioned within the emerging paradigm of intent-aware AI systems. Unlike conventional generative tools that output visuals devoid of spatial logic, the proposed system uses RAG to anchor AI suggestions in structured architectural data. This strategic orientation offers more than just visual variation—it enables context-specific reasoning and feedback, contributing directly to spatial decision-making and performance alignment.

In summary, this study seeks to redefine the architectural design pipeline by bridging the gap between inspiration and execution. It offers a model where AI not only generates but interprets, refines, and collaborates—positioning VR as an embodied design environment capable of hosting both ideation and specification tasks. In doing so, it responds to the most urgent demands in AI–VR research: interactivity, adaptability, explainability, and design relevance.

3 Hypothesis

This research hypothesizes that the integration of AI and VR platforms can transform small-scale architectural workflows by enabling designers to collaborate with AI agents in immersive environments. The process enhances real-time ideation, spatial scale awareness, and the generation of actionable deliverables, such as project specification books.

4 Methods

This research follows a practice-based computational design approach, focusing on developing and evaluating a system that integrates AI-driven image generation and modification with immersive VR interactions. The study consists of several experimental modules, including self-exploration of generative AI tools for panoramic image generation, AI-assisted early sketching and thematic exploration, material selection and specification using AI-driven recommendations, technical development of an AI-powered backend and VR-based frontend, and user testing to evaluate workflow efficiency and interaction quality. Together, these modules contribute to a deeper understanding of how AI and VR can be effectively integrated to enhance architectural design workflows.

4.1 Self-Exploration of Generative AI for Panoramic Image Generation

Before developing the AI-VR system, an exploratory phase was conducted to evaluate various generative AI tools for panoramic image generation. This self-exploration aimed to assess their architectural relevance, control mechanisms, and image fidelity.



Figure 1: Panoramic image of my bedroom as input

I used a panoramic image of my bedroom captured using an Insta360 panoramic camera as input to different tools for evaluation. The tools evaluated included Skybox AI, an online panoramic generation platform, Stable Diffusion with image-to-image generation, ControlNet and inpainting methods, and LLaMA 3.2 Vision Model for AI-assisted description and advisory tasks. Each tool was tested for its ability to generate, modify, and refine panoramic images within an architectural context, helping to establish a structured AI workflow for VR-integrated design ideation.

4.1.1 Experiment with Skybox AI for Panoramic Image Generation

I first tested an online panoramic image generation tool called Skybox AI, which allows users to generate 360-degree images by selecting preset styles and providing text-based descriptions. The platform offers a remix function that enables users to adjust the style details while maintaining spatial consistency.

Testing Process I uploaded the panoramic image of my bedroom as the base image. The input prompt was: "Interior of a bedroom with modern decoration, a bed next to the window, a desk with desktops on it, dining

tables and furniture, etc." Skybox AI generated high-quality panoramic images with accurate room structure and spatial arrangement. The remix function provided different style variations, allowing for diverse aesthetic explorations.



Figure 2: Example images of panoramic generation from Skybox AI in comparison to input(first row), from left to right are panoramic image, perspective 1, perspective 2.

Observations and Limitations Skybox AI produced high-quality panoramic images with minimal input, making it an efficient tool for quick design exploration. The remix function successfully maintained spatial relationships while generating stylistic variations, which proved valuable for early-stage ideation. However, the tool lacked functionality for minor modifications to specific elements, limiting its ability to refine designs iteratively. Additionally, it did not support object-specific changes, requiring users to regenerate the entire scene for any modification. While useful for initial ideation, its constraints make it less effective for detailed design refinements that require precise adjustments.

4.1.2 Initial Test with Stable Diffusion for Panoramic Image Generation

To explore other alternatives, I performed tests using Stable Diffusion image-to-image generation. In this exploration, I used the same panoramic image as input.

Image-to-Image Generation For the first experiment, I provided the original panoramic photo as input and prompted the AI with "a bedroom with modern decoration". The results showed that while certain spatial characteristics such as walls and ceilings were preserved, the AI completely transformed the objects within the scene: The dining table turned into a bed. A floor lamp's light became a window.Additionally in a 360 view setting, there is an obvious discontinuity of panoramic image edge.



Figure 3: Example images of panoramic generation from stable diffusion in comparison to input(first row), from left to right are panoramic image, perspective 1, perspective 2.

This indicated that while the model could maintain panoramic structure, its object persistence was inconsistent. The AI reinterpreted existing elements too far from existing image, making it unsuitable for faithful design refinements.

Image-to-Image with ControlNet (Canny Edge Detection) To improve alignment with the original image, I tested Stable Diffusion with ControlNet's Canny Edge detection. Using the same input panoramic image of my room and the prompt "a bedroom with modern decoration", the AI-generated scene successfully followed the overall spatial structure of the original image. However, the details became messy, with certain objects appearing distorted or misplaced. Even with the use of a 360-degree image LoRA model, the generated panorama exhibited obvious discontinuities at the edges, breaking the seamless immersive experience.



Figure 4: Example images of panoramic generation from stable diffusion in comparison to input(first row), from left to right are panoramic image, perspective 1, perspective 2. Third row image generated with LoRA to help persist panoramic quality and reduce discontinuous edge.

These results indicate that while ControlNet helps maintain spatial integrity, it struggles to ensure cohesive detail refinement in panoramic settings. This highlights the challenge of adapting AI-generated images for 360-degree applications, where maintaining seamless transitions between edges is crucial. Further refinement through specialized panoramic model training or post-processing techniques is needed to improve the usability of AI-generated panoramas in architectural workflows.

4.1.3 Experiment with Stable Diffusion Inpainting

The inpainting approach proved to be an effective method for making step-by-step refinements while maintaining the panoramic integrity of the image. Unlike full-image generations, inpainting allowed for targeted modifications, ensuring that only specific objects were altered without affecting the overall composition. This provided a higher level of control over object-specific modifications, making it a promising technique for iterative design processes.



Figure 5: Example images of panoramic generation from stable diffusion in comparison to input(first row), from left to right are panoramic image, perspective 1, perspective 3.

However, certain inconsistencies emerged in maintaining panoramic coherence. Elements within the inpainted areas occasionally misaligned with their surroundings, disrupting the seamless quality of the 360-degree image. This issue was particularly noticeable in areas with complex geometry or significant lighting variations, indicating the need for further optimization in panorama-aware AI models to enhance spatial continuity.

4.1.4 AI-Assisted Design Advisory Using LLaMA 3.2 Vision Model

To evaluate AI's potential as a design assistant, I tested Meta's LLaMA 3.2 11B Vision model as an AI-driven advisor for my bedroom renovation.

Due to GPU limitations, I deployed a LLaMA Stack server with used Together.AI's cloud AI model service. I used Meta LLaMA's official example llamastack application: "Interior Design Assistant", which provides an AI-powered agent capable of maintaining a consistent conversation with the user about interior design preferences. This app facilitates iterative design discussions, allowing users to input room images and receive AI-generated descriptive analysis and refinement suggestions.

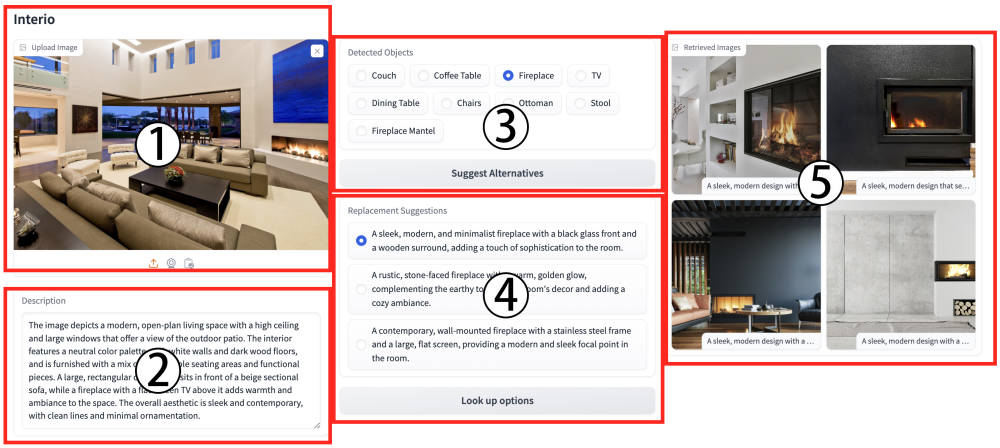


Figure 6: The interior design assistant app workflow: 1.User upload image; 2.The Llama AI agent gives a description ofthe image; 3.The AI agent list objects in the image, and user picks one for further suggestions on it; 4.The AI agent provide alternative descriptions about the selected object, and user picks from them; 5.The AI agent use RAG to retrieve example reference images that are related to the description.

Following the workflow of the app, I uploaded the original panoramic image of my bedroom, and the LLaMA 3.2 model processed it to generate a detailed textual description of the room's layout, furniture. I could see the output of the descriprion from the backend logs.

```
Running on local URL:  http://0.0.0.0:8003

To create a public link, set `share=True` in `launch()`.
Error parsing JSON output: The image depicts a room with a warm and inviting atmosphere, featuring a mix of modern and tra
ditional elements. The room is painted a soft yellow color, which complements the light-colored wood floors and creates a
sense of warmth and coziness.

The room is furnished with a variety of items, including a desk with a computer monitor and keyboard, a chair, a table wit
h a vase of flowers, a closet with clothes hanging inside, a bed with a white comforter, and a dresser with drawers. The f
urniture is arranged in a way that creates a sense of balance and harmony, with each piece complementing the others.

The room also features several decorative elements, including a lamp on the desk, a plant on the table, and a few pieces o
f artwork on the walls. The overall effect is one of comfort and relaxation, making the room feel like a welcoming and inv
iting space.

The room's architecture is characterized by its curved walls and ceiling, which create a sense of flow and continuity. The
 use of natural light from the window adds to the sense of brightness and airiness, making the room feel larger than it ac
tually is.

Overall, the room is a cozy and inviting space that combines modern and traditional elements in a way that creates a sense
 of balance and harmony. The use of warm colors, natural light, and comfortable furniture creates a welcoming atmosphere t
hat makes the room feel like a relaxing retreat.

**Furniture Items:**

* Desk
* Chair
* Table
* Closet
* Bed
* Dresser
* Lamp
* Plant
* Artwork
* Computer monitor
* Keyboard
* Vase
* Flowers
* White comforter
* Drawers
```

Figure 7: LLaMA 3.2 model generated detailed textual description of the room's layout, furniture object lists.

However, the app encountered a failure in structured output formatting: From the backend code we can tell that the backend expected AI output in a structured JSON format, but LLaMA 3.2 returned unstructured free text. Due to this issue, the UI could not interpret the response, preventing further interaction.

```
assert (
    self.agent_id is not None
), "Agent not initialized, call initialize() first"
text = textwrap.dedent(
    """
    Analyze the image to provide a 4 sentence description of the
    architecture and furniture items present in it.
    Description should include architectural style, design, color,
    patterns, textures and other prominent details.
    Examples for furniture items include (but not limited to): couch,
    coffee table, fireplace, etc

    Return results in the following format:
    {
        "description": 4 sentence architectural description of the
        image,
        "items": list of furniture items present in the image
    }

    Remember to only list furniture items you see in the image. Just
    suggest item names without any additional text or explanations.
    For eg. "Couch" instead of "grey sectional couch"

    Please return as suggested format, Do not return any other text or
    explanations.
    """
)
```

Figure 8: Backend code for generating image descriptions, which speculated the format of output concisely.

This experiment highlighted the potential of AI-driven brainstorming in interior design while also revealing LLaMA 3.2's limitations in instruction adherence, necessitating improved prompt engineering and error-handling mechanisms in future iterations.

4.1.5 Summary of Findings from Self-Exploration

Two-Step AI Workflow Based on the findings, I structured my AI-driven generation workflow into two steps: Initial Style Exploration: Use Skybox AI or Stable Diffusion's image-to-image generation to create an initial panoramic style transformation. Iterative Refinement: Apply Stable Diffusion's inpainting method for controlled object modifications. LLaMA 3.2 will be integrated as an advisory AI at each stage to assist with prompt refinement and output analysis.

Key Limitations and Next Steps The panoramic coherence of Stable Diffusion needs improvement, which requires LoRA fine-tuning. AI-generated object transformations remain unpredictable, limiting precise control in design workflows. LLaMA 3.2 struggles with instruction-following, requiring better prompt engineering for structured outputs. This self-exploration phase confirmed that while current AI tools show promise, they require custom adaptations to be effectively used in interactive VR-driven architectural workflows.

4.2 Technical Details of System Development

To explore the research question—how the convergence of AI with VR could impact the architectural design pipeline—I developed a system titled "MerGen" as a foundational prototype platform. It has the nature of both immersion and generation. It was built for supporting several experimental modules to serve as a functional platform for immersive, multimodal interaction and AI-assisted environmental design. The system is architected as a client-server model composed of a Unity-based frontend optimized for VR interaction, and a containerized backend composed of modular AI and image processing services. This end-to-end infrastructure enables real-time voice-driven commands, image-based scene analysis, panorama generation, and localized scene refinement.



Figure 9: MerGen-a hybrid of immersion and generation

The overall objective of the system is to facilitate seamless collaboration between user and AI within a spatial design workflow. The user engages with the system through natural voice commands and intuitive hand gestures using the Meta Quest device. The frontend is responsible for speech processing, scene rendering, and interaction capture, while the backend handles semantic reasoning, image understanding, scene synthesis, and WebSocket-based state synchronization.

Key interactions in the system begin with the user's voice input, which is transcribed and semantically analyzed in the Unity frontend via Wit.ai. The extracted intent, along with any image or mask data, is forwarded to the Django-based backend, where it is routed to the appropriate microservice—whether it be for LLaMA-based conversation and image analysis, panoramic image generation, or mask-guided inpainting. All modules share access to a synchronized image directory that links Unity's asset folder with the backend's working path, ensuring consistent visual updates.

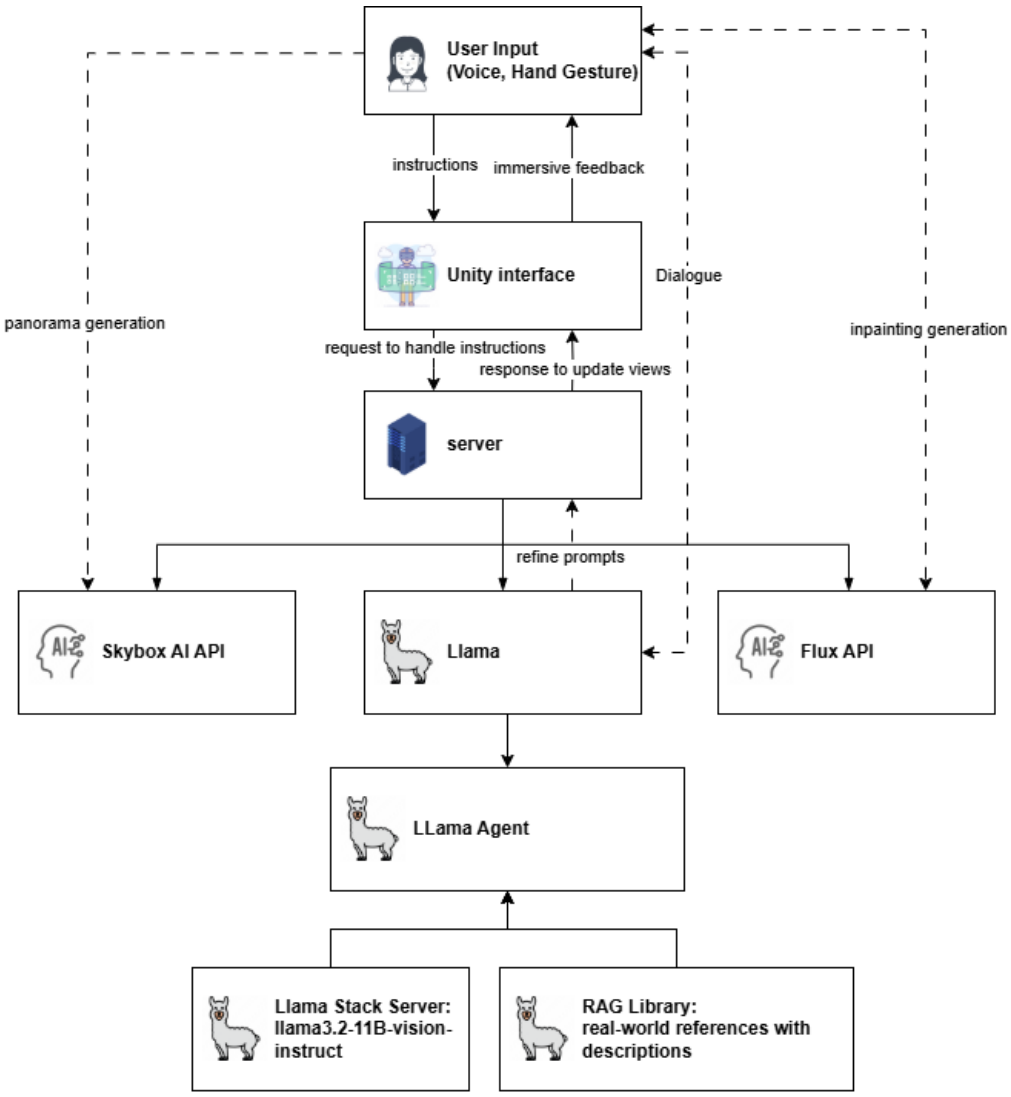


Figure 10: MerGen architecture diagram

This modular system architecture ensures extensibility and robustness while offering low-latency feedback loops necessary for immersive user experience. The following two subsections detail the specific implementation of the backend AI framework and the Unity-based frontend environment.

4.2.1 Unity-Based VR Frontend

The frontend component of the MerGen system is developed using the Unity engine, targeting the Meta Quest platform as the primary runtime environment. All development is aligned with the constraints of VR hand-tracking interaction, OpenXR input mappings, and real-time rendering requirements appropriate for standalone headsets. The frontend system supports immersive scene visualization, natural language command recognition, and spatial editing through direct hand gestures. These core functionalities are implemented through a combination of Unity scripting, XR input modules, custom shaders, RestAPI and WebSocket-based real-time communication with the backend.

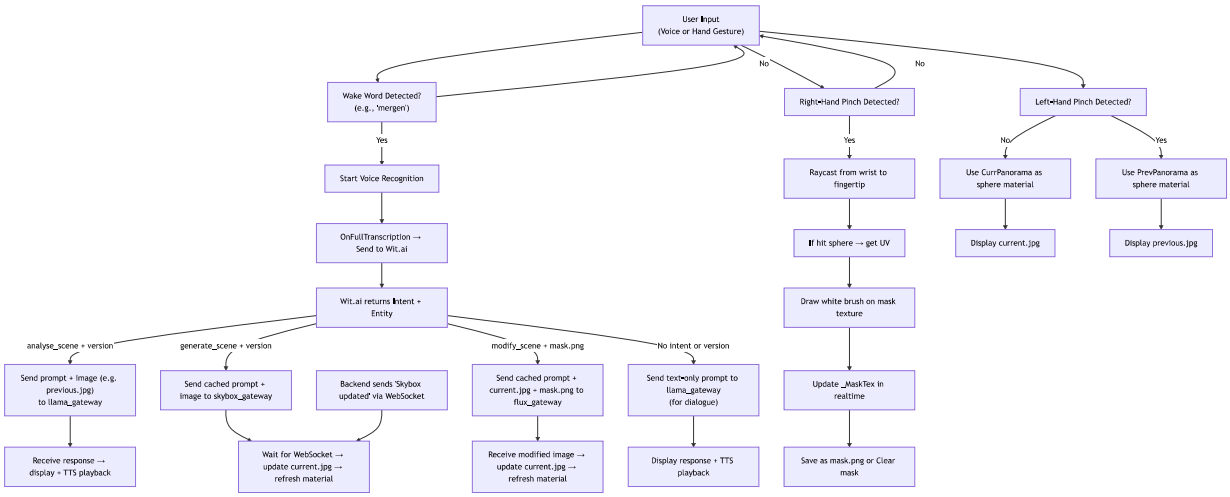


Figure 11: User input processing flow in the Unity-based frontend, showing voice interaction, gesture detection, intent routing, and image update logic.

Panoramic Visualization At the center of the immersive environment is a panoramic rendering system based on a spherical mesh. To avoid distortions commonly caused by default UV mappings, the mesh uses customized UVs specifically designed for equirectangular image projection. This sphere serves as the canvas onto which panoramic images are rendered, giving the user a complete 360-degree visual environment.

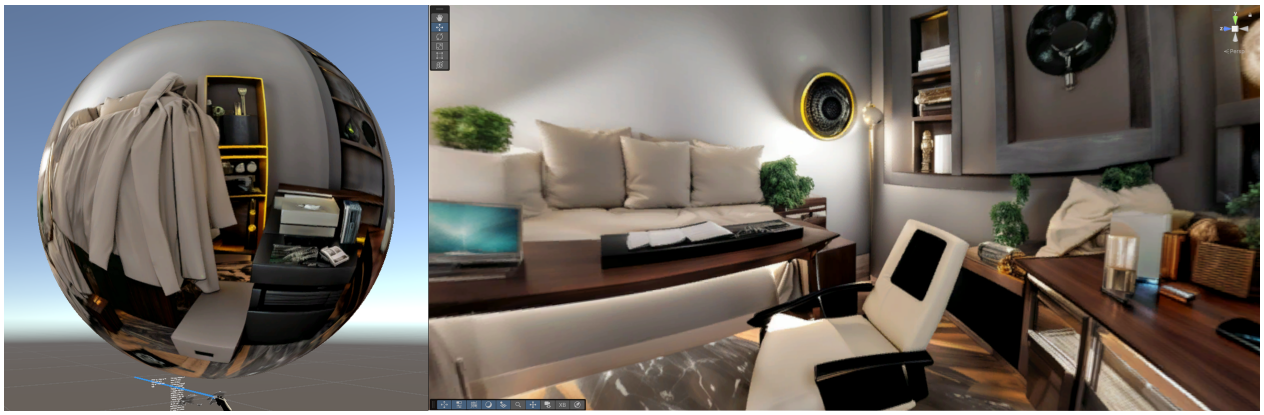


Figure 12: Spherical mesh with customized UVs specifically designed for equirectangular image projection. Left: sphere: right: view from inside sphere

The sphere's material dynamically switches between two texture states: CurrPanorama (which displays current.jpg) and PrevPanorama (which displays previous.jpg). These textures are shared between Unity and the backend via a mounted folder (Assets/Material) and are updated by backend operations. Each time a new panorama image is generated, the backend archives the current one as previous.jpg and stores the new image as current.jpg. This allows temporal versioning and enables comparative visualization.

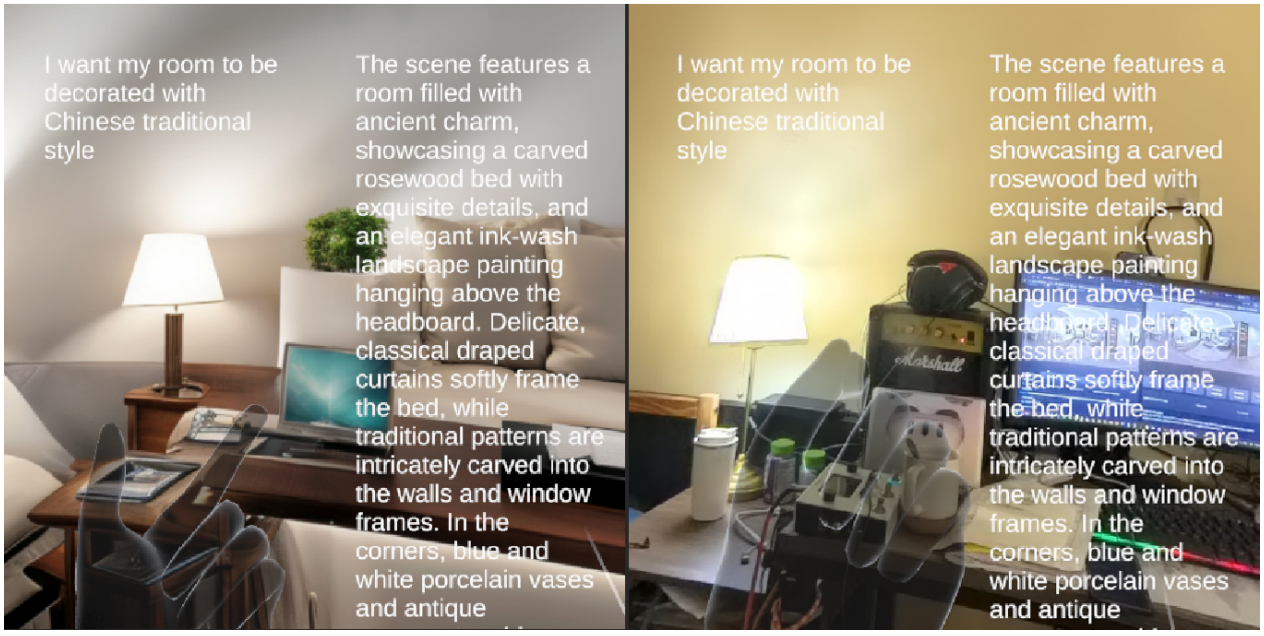


Figure 13: Left pinch gesture will switch scenes. Left: while not pinched it shows the latest iteration, Right: while pinched, it shows the last iteration.

A left-hand pinch gesture, detected through the XR Interaction Toolkit's SelectAction, toggles the panoramic sphere's material between current and previous versions. This physical gesture allows the user to instantly compare iterations of the scene and assess changes without leaving the immersive environment.

Speech Interaction and Intent Recognition Speech control is implemented via the Meta Voice SDK, leveraging Wit.ai for natural language understanding. The interaction sequence begins with wake word detection—`"mergen"`, `"hi mergen"`, or similar variations. Upon recognition, the system emits an activation sound, and the following spoken sentence is treated as a command. After processing, a deactivation sound plays, and the system pauses listening until the wake word is triggered again.

Speech transcription is managed by the `VoiceManager` script in Unity, which listens for partial and full speech events. Partial transcriptions are rendered in real time using a TextMeshPro UI element, while full transcriptions are dispatched to the backend as intent-bearing commands.

Wit.ai is configured with six custom intents:

- `wake_word`: indicates the system is activated.
- `analyse_scene`: triggers image analysis requests.
- `generate_scene`: triggers new image generation based on prior dialogue.
- `modify_scene`: initiates inpainting based on a mask.
- `show_example`: triggers retrieval-augmented generation (RAG) to display real-world product examples based on the latest LLaMA-generated description.
- `output_specification`: instructs the LLaMA agent to summarize the current design session and generate a specification document, which is saved as a timestamped file in the backend for archival.

The version entity (`version:version`) specifies the base panoramic image to be used in scene generation or analysis. Values such as `"new"`, `"old"`, and `"original"` are mapped to `current.jpg`, `previous.jpg`, and `original.jpg`, respectively. When both an intent and a version are detected, the Unity frontend dispatches a structured request to the backend. For instance, if the `analyse_scene` intent and the version `"old"` are detected, Unity sends both the spoken input and the `previous.jpg` image to the backend for image-text analysis.

If no intent is detected, the raw speech transcription is forwarded to the LLaMA backend as a general conversation input. This fallback ensures that user input is always processed, even outside of predefined interaction patterns.

Before executing a `generate_scene`, `modify_scene`, or `show_example` operation, the system caches the most recent LLaMA response. This response is used as the prompt for the respective generation or retrieval task, maintaining semantic continuity between conversation and visual outputs.

Finally, the system uses the Meta Voice SDK's `TTSSpeaker` module to deliver backend responses via speech. All responses—whether they are chat replies, generation confirmations, or specifications—are read aloud to the user. Longer outputs are automatically chunked and streamed in sequence to ensure smooth playback and maintain immersion.

In-Scene Editing with Mask Painting In addition to voice control, the frontend supports spatial editing through a real-time mask painting interface. This allows users to select and modify specific regions of the panoramic image by gesturing directly within the scene.

When the user performs a right-hand pinch, the XR SelectAction event triggers a raycast from the right wrist toward the index fingertip. This raycast is visualized with a LineRenderer, which only appears during active pinch events to avoid visual clutter.

If the ray intersects with the panoramic sphere, the hit point's UV coordinates are calculated. A white circular brush is then applied to a Texture2D mask at that UV location. This mask texture is initialized in memory as a transparent black image and incrementally updated in real time. It is bound to the material shader through a _MaskTex property, allowing immediate visual feedback inside the headset.

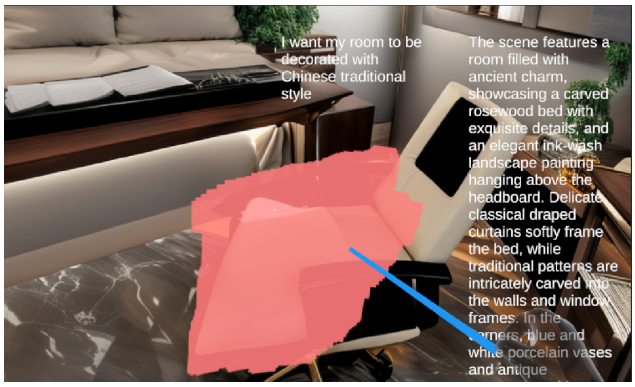


Figure 14: Right pinch gesture will show a ray from user's hand acting as the inpainting pen.

Users can save the mask with head noding guesture to persist the drawing as mask.png in the Assets/Material folder. A shaking head gesture will reset the mask to all-black, effectively erasing all edits. This functionality is particularly useful for inpainting scenarios, where users want to refine only specific parts of the current panorama image.

Backend Communication All Unity-side operations rely on structured backend interactions facilitated via RESTful APIs exposed by the Django gateway service. Depending on semantic intents recognized by Wit.ai, the Unity frontend routes specific user commands to their respective backend endpoints, enabling seamless multimodal interactions between the frontend and backend.

Analyse Scene Intent:

Upon recognizing the analyse_scene intent along with an accompanying version entity (such as "new", "old", or "original"), the Unity frontend identifies the appropriate panoramic image file (current.jpg, previous.jpg, or original.jpg). It then sends both the transcribed user command and the selected image file to the Django backend's llama_gateway API. This request triggers the ida-llama-agent service, which invokes the LLaMA model to perform image analysis combined with semantic text interpretation. Subsequently, the model's textual analysis and recommendations are returned to the frontend, where they are visually displayed and audibly read back to the user through the Text-to-Speech (TTS) system.

Generate Scene Intent:

When the "generate_scene" intent is detected, the frontend again validates the presence of a specified version entity to determine the base reference image for generating a new panoramic scene. Utilizing the most recent response text received from the LLaMA model as the descriptive prompt, Unity submits this data to the backend's panorama generation endpoint (pano-gen/generate_with_image/). The backend packages this request, including the image and the prompt, into a structured JSON payload and sends it to the Blockade Labs Skybox AI service for panoramic image generation. As Skybox AI generates images asynchronously, the backend registers a webhook to receive updates upon completion. When notified, the backend updates the images (current.jpg becomes the new image and the existing image is archived as previous.jpg) and sends a real-time WebSocket notification to Unity. Upon receiving this notification, Unity dynamically reloads the updated panorama, providing immediate visual feedback within the VR environment.

Modify Scene Intent:

Upon recognizing the modify_scene intent, the system presumes the user has already created a mask image (mask.png) by performing gesture-based painting within the VR environment to indicate regions requiring refinement. The Unity frontend retrieves this mask image, the current panoramic image (current.jpg), and the latest text response from the LLaMA model. These files and prompts are sent to the Django backend's inpainting endpoint (pano-gen/inpainting/). The backend then sends this data to the Flux Inpaint API (Segmind) to perform targeted modifications within the masked regions. Once complete, the modified image is saved as the new current.jpg, and a real-time WebSocket notification updates the VR frontend, providing instantaneous visualization of the refined image.

Show Example Intent:

When the show_example intent is triggered, the frontend submits the most recent descriptive response from the LLaMA model to the backend, initiating a Retrieval-Augmented Generation (RAG) operation. Leveraging a pre-built knowledge base (comprising IKEA furniture product descriptions and associated metadata, including prices and image URIs), the backend retrieves three best-matching items. These items—structured as JSON arrays containing image paths, prices, and concise descriptions—are sent back to Unity. The frontend visually presents the highest-ranking result prominently within the VR environment, displaying the product image, price, and summarizing description. Users dissatisfied with the suggested product can perform a head-shaking gesture to dismiss the displayed recommendation, reverting the visualization back to the default state.

Output Specification Intent: When the output_specification intent is triggered, the Unity frontend sends a fixed prompt—"Based on our earlier discussion, write a specification as a summary about all of the design decisions we made and why, including styles, materials, prices."—to the backend's llama_gateway endpoint. This request is forwarded to the ida-llama-agent service, which uses the session's ongoing conversational memory to generate a comprehensive design specification. The agent's response is saved to a timestamped .txt file in the logs directory. This ensures that each generated specification is archived for later reference or export.

Fallback (No Intent Detected):

If user speech input does not match any predefined Wit.ai intents or lacks the necessary entities, the Unity frontend transmits only the raw text transcription directly to the backend's LLaMA conversational service. The LLaMA model subsequently processes this input as a general dialogue query, returning a textual

response that is both visually and audibly conveyed to the user. This fallback mechanism ensures consistent and uninterrupted user-AI interactions within the VR design environment.

4.2.2 Backend AI Framework

The backend of the MerGen system is designed as a modular, containerized microservice architecture capable of supporting multimodal interaction through image understanding, real-time scene generation, and modification. The backend services receive structured HTTP requests consisting of parsed user text and, when applicable, associated images or masks.

The backend is implemented using Docker Compose, which orchestrates multiple services and facilitates isolated development, deployment, and scaling. All services communicate over an internal bridge network named mergen-network, using REST APIs and WebSocket protocols.

Core Gateway Architecture At the heart of the backend is the server container, a Django-based project that functions as the main API gateway. It exposes HTTP and WebSocket interfaces to the Unity frontend and provides internal routing logic to handle various multimodal intents. The Django project is modularized into three independent applications:

- **llama_gateway:** Responsible for general dialogue, image analysis, and multimodal interaction with the LLaMA model.
- **skybox_gateway:** Responsible for full-scene panorama generation via external APIs.
- **flux_gateway:** Responsible for inpainting-based image editing based on user-painted masks.

All front-end interactions—whether text-only requests, image analysis commands, or editing actions—are funneled through this unified API layer.

LLaMA-Based Conversational and Visual Intelligence The `llama_gateway` application manages all requests related to conversational AI and multimodal image analysis. When the Unity frontend detects an analyse_scene or general dialogue command, the gateway determines whether the request is text-only or includes a reference to an image (via the version entity). The designated image—either current.jpg, previous.jpg, or original.jpg—is located in the shared volume /app/Material, loaded into memory, and optionally compressed.When the Unity frontend detects a show_example intent, the gateway is gonna ask the agent to trigger a RAG tool use for the next response.

These requests are passed to the ida-llama-agent container, which is itself a Django-based backend service. This container handles the lifecycle of an intelligent agent instance built on the LLaMA model and exposes REST API endpoints for receiving multimodal input and returning generated responses. The agent is initialized at runtime using the llama_stack, which connects to a separate container llama-stack-server.

Upon startup, ida-llama-agent instantiates a custom subclass called InterioAgent. This agent is configured with a persistent identity using the Together AI model meta-llama/Llama-3.2-11B-Vision-Instruct-Turbo and is initialized with expert-level interior design instructions. These include domain-specific constraints such as

the use of spatial vocabulary, furniture styles, materiality, and lighting considerations. This ensures that all replies are grounded in the architectural design domain. Sampling is configured to use greedy decoding for deterministic outputs.

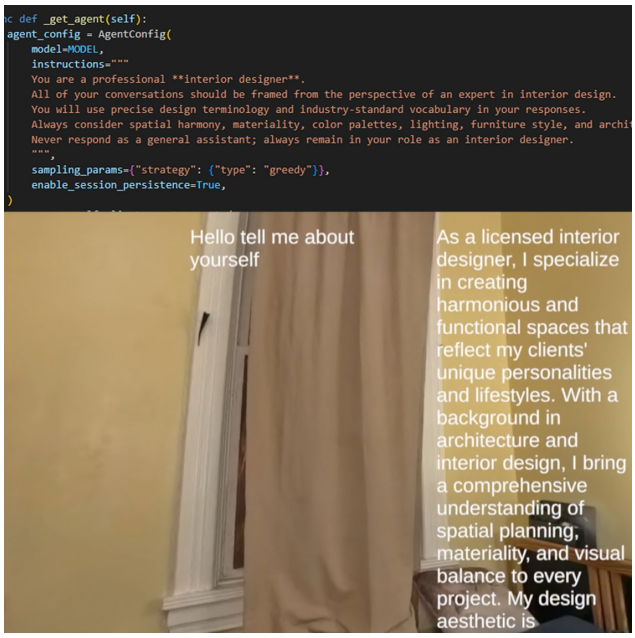


Figure 15: Llama Agent is created as an interior designer. Top: code for initialization. Bottom: dialogue in Unity, asking the agent about its role.

A memory-augmented context environment is also created at initialization. The system loads .txt documents from a specified document_dir, embeds them using the all-MiniLM-L6-v2 model, and stores them in a registered vector database under a live memory bank ID (interio_bank). This setup enables retrieval-augmented generation (RAG), allowing the agent to answer user queries with contextual awareness and prior descriptive grounding.

For every incoming user query, a new session is created using a UUID-based identifier, enabling persistent dialogue tracking. The query is transformed into a structured multimodal message object, including both the spoken prompt and optionally an attached image (converted into a base64 data_url). This message is streamed to the agent using a turn-based API, and output is accumulated token by token. Once the response is finalized (turn_complete), the agent's reply is sent back to the Unity frontend for rendering and speech synthesis.

The llama-stack-server container functions as a lightweight proxy service that mediates all model inference requests. It forwards requests from ida-llama-agent to the Together AI cloud model API and returns the streaming output. This modular design abstracts away hardware dependency, enabling powerful vision-language reasoning without requiring local GPU resources. It also facilitates centralized configuration of model selection and prompt behavior without coupling to frontend or application logic.

Incoming requests from the Unity frontend are handled according to the intent recognized by Wit.ai, resulting in distinct processing pathways:

1. **General Dialogue (No Intent Detected):** When no explicit intent is detected by Wit.ai, the user's spoken input is treated solely as conversational text. The Unity frontend forwards this plain-text query directly to the ida-llama-agent container. Here, the agent processes the text query using the LLaMA model configured via the llama_stack, generating contextually relevant conversational responses. Once generated, the reply is sent back to the Unity frontend, where it is visually displayed and read aloud via Text-to-Speech (TTS).

2. **Analyse Scene Intent:** If the Unity frontend identifies the `analyse_scene` intent and confirms the presence of a version entity (e.g., "new", "old", or "original"), the corresponding panoramic image (current.jpg, previous.jpg, or original.jpg) is selected from a shared volume (/app/Material). The frontend submits both the user's spoken query and the specified image to the backend. The ida-llama-agent then invokes the multimodal LLaMA model through the llama_stack inference service, performing visual analysis combined with textual semantic interpretation. The agent's detailed analysis and recommendations, derived from both textual and visual inputs, are subsequently returned to the frontend as a coherent response.

3. **Show Example Intent (Retrieval-Augmented Generation):** When the frontend detects the show_example intent, it triggers a specialized Retrieval-Augmented Generation (RAG) process. Instead of directly using the user's current query, the system utilizes the last response provided by the LLaMA agent as a descriptive context. Specifically, the previously generated conversational reply serves as the search query (description) for retrieving matching items from an internal knowledge repository:

   The backend agent (ida-llama-agent) maintains a vector-embedded knowledge base, built from detailed text documents describing various furniture items, each including structured data such as image paths and pricing.

   The retrieve_images method within the agent executes the RAG process, querying the vector database to identify and retrieve three best-matching items based on semantic similarity to the provided description.

   The resulting retrieval output strictly adheres to a predefined JSON structure, for instance:

```
[
  {
    "image": "001.jpeg",
    "price": "$100",
    "description": "A beautiful table with a modern design."
  },
  {
    "image": "009.jpeg",
    "price": "$120",
    "description": "A stylish chair with a vintage look."
  },
  {
```

```
    "image": "006.jpeg",
    "price": "$160",
    "description": "A cozy sofa with a contemporary style."
  }
]
```

Once received by the frontend, the top-matching retrieved item is visually presented within the VR environment, displaying the item's image, price, and descriptive summary. Users can subsequently dismiss the displayed recommendation through a head-shaking gesture, returning the interface to its default visual state.

Panorama Image Generation The `skybox_gateway` is triggered when the frontend detects the `generate_scene` intent. The Unity application sends the prompt and a `version` value identifying the reference image to be used as context. The backend loads the appropriate image (e.g., `current.jpg` or `original.jpg`) and encodes it in base64.

The image, along with the user prompt, is packaged into a JSON payload and sent to the Blockade Labs Skybox AI API. The payload includes:

- `skybox_style_id`: predefined style identifier.
- `prompt`: the text prompt obtained from the previous LLaMA response.
- `control_image`: the base64-encoded panorama image.
- `control_model`: set explicitly to ‘‘`remix`’’.
- `webhook_url`: auto-generated by the backend via Ngrok.

Due to the asynchronous nature of the Skybox API, the backend registers a webhook endpoint at `/api/pano-gen/webhook/` to receive progress updates. When a generation task completes, the backend performs the following steps:

1. Archive the existing `current.jpg` as `previous.jpg`.
2. Download the newly generated image from the provided `file_url`.
3. Save the result as the new `current.jpg`.
4. Send a WebSocket notification through `/ws/skybox-updates/` to update the Unity frontend.

To support webhook callbacks in a local development environment, the backend uses an `ngrok` container to expose public URLs. It continuously polls the Ngrok API at:

```
http://ngrok:4040/api/tunnels
```

until a valid HTTPS URL is found, which is then injected into the external API request.

Inpainting-Based Image Editing The `flux_gateway` processes requests triggered by the `modify_scene` intent, which is recognized through voice commands. Prior to this, the user may have painted a mask over undesired panoramic regions using a right-hand gesture. These masks are stored locally as `mask.png` within the `Assets/Material` directory.

When the intent is triggered, the Unity frontend sends a structured request to the backend, which includes:

- `image`: the current panoramic image (`current.jpg`).
- `mask`: the binary user-edited mask image (`mask.png`).
- `prompt`: the most recent response generated by LLaMA.

The backend reads the image and mask, then assembles a POST request to Segmind's Flux Inpaint API. After receiving the updated image, the backend replaces the previous `current.jpg` file. This process enables precise, targeted image refinement without requiring complete scene regeneration.

Deployment Configuration and Integration All backend services are orchestrated via `docker compose`, which defines and connects the following containers:

- `server`: Django-based API gateway exposing REST and WebSocket endpoints.
- `ida-llama-agent`: manages multimodal AI agents and vector-based RAG retrieval.
- `llama-stack-server`: a lightweight proxy to Together AI's cloud-based LLaMA inference API.
- `ngrok`: provides secure public tunneling for webhook accessibility.

All services use a shared `.env` configuration file. The `server` container mounts a volume from `../Frontend/Assets/Mat` to `/app/Material` to enable shared access to panoramic images and masks.

Real-time frontend updates are enabled via Django Channels and Redis, supporting WebSocket-based notifications. All generated image files are versioned and replaced in-place, maintaining scene history and ensuring compatibility with prior states throughout the iterative design process.

This backend infrastructure provides a scalable, modular, and developer-friendly environment to support MerGen's immersive design pipeline powered by multimodal interaction and cloud-based reasoning.

4.2.3

The experimental evaluation of MerGen was structured around two integrated design modules reflecting distinct stages of the architectural workflow: early-stage spatial ideation (chapter 4.3) and material specification(chapter 4.4). In the first module, I used natural language dialogue and gesture-based selection to generate and refine panoramic spatial scenes in real time. This module tested the system's ability to semantically translate abstract design prompts, support interactive exploration of spatial alternatives, and enable intuitive iteration through localized inpainting. The second module focused on bridging conceptual design with implementation detail. Here, the AI agent provided stylistically aligned material recommendations using a retrieval-augmented generation (RAG) pipeline, allowing users to preview substitutions directly

within the immersive scene. These two modules were evaluated through five scenario-based tasks, simulating realistic residential and commercial design briefs, and collectively examined MerGen's potential to facilitate a continuous, immersive, and designer-centered workflow.

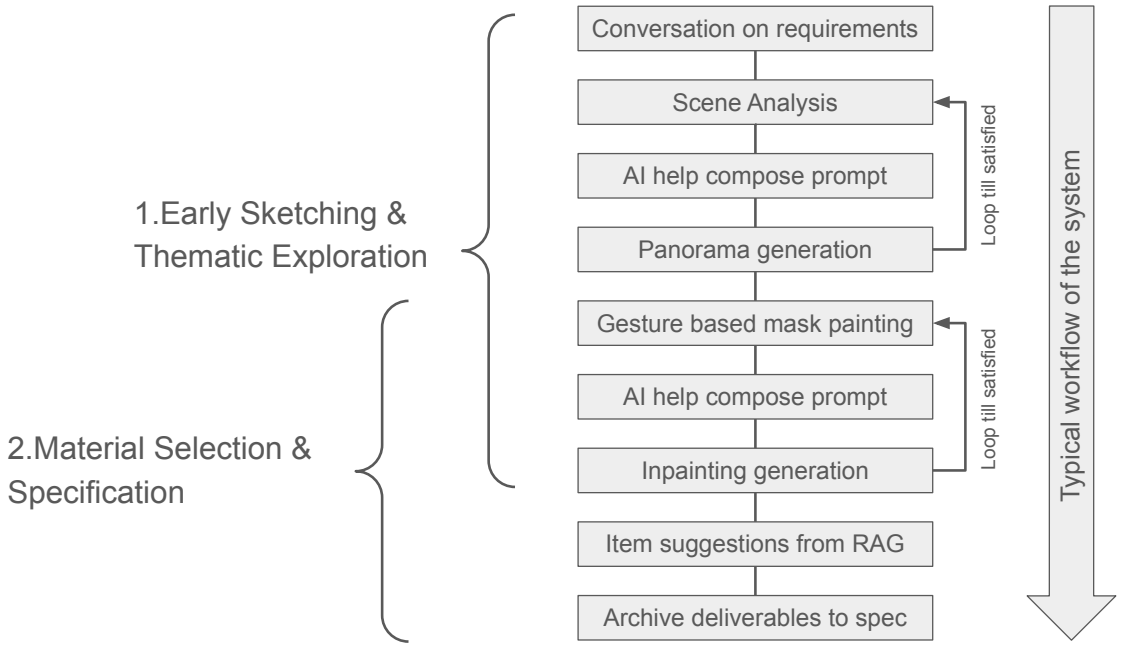


Figure 16: Typical workflow of the system and 2 experiment modules.

4.3 Early Sketching and Thematic Exploration

4.3.1 Research Goals and Rationale

Early-stage architectural design is characterized by rapid ideation, extensive thematic exploration, and iterative refinement of spatial concepts. During this critical phase, designers often navigate complex trade-offs between aesthetic coherence, functional requirements, spatial constraints, and client-specific preferences. Conventional workflows, while flexible, typically lack direct integration between conceptual ideation, visual exploration, and iterative modification—particularly within immersive or intuitive spatial contexts. This gap can lead to inefficiencies, misaligned design outcomes, and limited creative exploration [32, 13].

The MerGen prototype was specifically developed to address these challenges by integrating conversational artificial intelligence (AI), real-time panoramic scene generation, and gesture-based spatial editing within an immersive virtual reality (VR) environment. By combining natural language interaction, multimodal visualization, and direct manipulation capabilities, MerGen aims to support architects in efficiently navigating the complexities of early design ideation [35, 15].

In particular, the research presented here evaluates MerGen's effectiveness in three critical dimensions of early-stage design:

- **Semantic Translation of Abstract Intentions**: Assessing how effectively MerGen interprets designers' abstract descriptions, style references, and functional requirements, accurately translating these into coherent spatial concepts through dialogue-driven interactions [13].
- **Generative Exploration of Spatial Alternatives**: Evaluating the ability of MerGen to generate relevant and diverse panoramic visualizations aligned with explicitly stated design goals and stylistic intentions, thus facilitating meaningful thematic exploration and spatial experimentation [28, 3].
- **Interactive Iteration and Refinement**: Investigating MerGen's capabilities for intuitive and responsive design refinement, particularly through targeted mask-based inpainting, enabling designers to iteratively correct and adjust visual outcomes without full scene regeneration [35].

Furthermore, this evaluation explicitly addresses several critical issues raised by recent industry surveys, such as those by the American Institute of Architects (AIA), including:

- **Reliability of AI-generated Outputs**: Ensuring that the generated visual and spatial outcomes reliably align with the initial intentions expressed by the designer and their clients [19].
- **Transparency and Explainability**: Assessing whether the rationale behind design suggestions and generative results provided by the conversational AI is sufficiently transparent, understandable, and actionable for professional designers [30].
- **Agency and Control in Human-AI Collaboration**: Exploring whether MerGen effectively supports designer agency, empowering architects rather than constraining their creative autonomy [15, 16].

To comprehensively evaluate these objectives, I designed a structured series of five diverse experimental scenarios, each representing realistic architectural contexts and client profiles. These scenarios included residential interiors (such as minimalist urban apartments and tranquil bedroom-office hybrids), collaborative living environments (e.g., multifunctional social spaces), and small-scale commercial refurbishments (an aging flower shop renovation). Each scenario was carefully crafted to challenge MerGen's semantic understanding, visual creativity, and interactive refinement capabilities within varied and realistic design briefs.

Collectively, these experiments aimed to thoroughly document MerGen's effectiveness in enhancing early-stage design processes, critically examining its strengths, limitations, and overall potential for integration within professional architectural workflows.

4.3.2 Experimental Workflow

To systematically evaluate MerGen's capabilities in supporting early-stage design ideation, I conducted each experimental scenario following a structured and consistent multi-step workflow. This workflow was specifically designed to leverage MerGen's multimodal capabilities—combining conversational dialogue, panoramic scene generation, and iterative visual refinements—to explore spatial concepts, address client requirements, and iteratively approach the intended design outcome. Each step was clearly defined and carefully documented to ensure repeatability, clarity, and detailed assessment of the AI-human collaborative process:

1. **Initial Dialogue and Client Briefing:**

    - At the outset of each scenario, I clearly articulated a detailed client profile, including demographic characteristics, lifestyle preferences, functional needs, and explicit spatial requirements.
    - This description was conveyed through structured natural language conversations with the LLaMA agent, ensuring comprehensive semantic context and enabling the AI to accurately interpret abstract stylistic and functional intentions.

2. **Spatial Analysis and Diagnostic Feedback:**

    - Following the initial briefing, the LLaMA agent analyzed the provided spatial input—typically a 360° panoramic image or a neutral white-box VR model.
    - Through conversational interaction, the agent identified and described existing spatial issues (e.g., clutter, poor lighting conditions, inefficient layouts), opportunities for enhancement, and relevant stylistic considerations based on initial client descriptions.
    - This provided a clear diagnostic foundation, framing subsequent discussions and design interventions.

3. **Collaborative Stylistic Exploration and Prompt Refinement:**

    - Building upon the initial analysis, I engaged the LLaMA agent in an iterative dialogue, systematically exploring and refining specific stylistic concepts, spatial arrangements, material choices, and visual themes.
    - Throughout this phase, the agent proactively offered domain-specific suggestions (e.g., suitable furniture types, appropriate lighting arrangements, material palettes), while I continually clarified or expanded upon these suggestions, ensuring alignment with the stated design objectives.
    - Through multiple iterations of dialogue-driven exchanges, the conversational process converged into a detailed, concrete prompt designed specifically for subsequent visual generation.

4. **Panoramic Visualization via AI Generation:**

    - Using the refined semantic prompt obtained from collaborative dialogue, I triggered a panoramic visualization generated by Skybox AI. This visualization offered an immersive, 360° perspective on the proposed spatial design.
    - Generated panoramic scenes served as initial spatial hypotheses, directly reflecting the thematic intent established through dialogue and offering tangible visualizations of previously abstract design goals.
    - Immediately following visualization, each generated image was carefully reviewed in terms of thematic coherence, spatial accuracy, stylistic consistency, and adherence to explicitly stated client preferences.

5. **Interactive Visual Refinement through Mask Painting and Inpainting:**

- Recognizing that initial AI-generated visualizations often required specific adjustments, I utilized MerGen's mask-based painting tool for precise, targeted visual modifications.
- By identifying and painting masks directly within the VR environment, I clearly demarcated visual regions needing improvement or stylistic corrections.
- These masked areas were subsequently refined using the Flux Inpainting API, integrating new visual elements based on explicit follow-up prompts collaboratively generated with the LLaMA agent.
- This iterative refinement step was essential for ensuring detailed alignment between the generated visualization and the intended design outcome, effectively correcting inaccuracies and further enhancing spatial coherence and thematic alignment.

6. **Iterative Conversational Feedback Loop:**

   - Throughout the experimental workflow, I continuously engaged in conversational feedback loops with the agent, critically assessing generated results, clearly communicating any discrepancies, and iteratively refining prompts for subsequent visual adjustments.
   - These repeated dialogue-driven refinements allowed for incremental improvements, closely approximating final spatial concepts through successive iterations.
   - This feedback loop was intentionally designed to mimic authentic architectural practice, providing realistic evaluation conditions for assessing MerGen's iterative flexibility, responsiveness, and conversational effectiveness.

By adhering strictly to this structured workflow across all scenarios, I ensured consistency in my methodological approach, enabling thorough documentation, detailed comparative analysis, and robust evaluation of MerGen's capabilities and limitations throughout the early stages of architectural design exploration.

4.3.3 Detailed Experimental Narratives

Experiment 1: Serene Bedroom Transformation **Client Brief:** A remote freelancer seeking a serene, multifunctional bedroom-office hybrid. The client required clear zoning for relaxation and productivity, emphasizing soft lighting, natural textures, and visual cleanliness to support both deep rest and focused solo work.

The experiment began with an initial conversational interaction with the LLaMA agent, clearly articulating the freelancer's unique requirements. The agent successfully identified the client's core desires, recommending distinct zoning strategies for rest and work, suggesting subtle divisions through furniture arrangements, lighting differentiation, and the strategic placement of calming natural textures (e.g., wood, woven fibers, and soft textiles). LLaMA also highlighted the importance of minimalist design principles, proposing neutral color palettes and limited decorative items to ensure visual tranquility and spatial clarity.



Figure 17: Existing panoramic image of the bedroom.

Following the initial conceptual dialogue, I prompted the agent to analyze an existing panoramic image of the bedroom, to which LLaMA responded accurately by identifying significant clutter and visual disorganization. LLaMA suggested a strategic decluttering plan alongside furniture and material recommendations that would enhance the sense of calm and productivity, effectively aligning with the client's brief.

With these insights, I iteratively refined the prompt with LLaMA, emphasizing minimalist aesthetics, soft neutral tones, and clearly delineated functional zones for work and rest. Despite careful semantic refinement, the first two generated panoramic visualizations did not fully meet expectations: the initial image presented an overly bright workspace that clashed with the relaxing atmosphere, while the second iteration exhibited awkward spatial divisions and lacked cohesive aesthetics. Notably, during the dialogue, a series of systematic interpretation errors occurred: for example, the agent repeatedly misunderstood certain terms, such as interpreting "bedroom" as "bathroom," and misinterpreting the phrase "the client also wants" as "decline also one." Such semantic errors disrupted the conversational fluency and delayed effective iterations.

After further detailed semantic clarifications, the third generated panoramic image provided a more suitable initial design base, aligning significantly closer to the desired aesthetic and functional intentions.



Figure 18: Tree initial panoramic generation of the serene multifunctional bedroom, with the last one accepted, despite window and door inaccuracies. Llama generated prompt used for the last image: “Generate an image of a serene and functional room with a minimalist aesthetic, featuring a comfortable seating area with a statement piece of furniture, a dedicated workspace with a desk and task lighting, and a peaceful sleeping area with a plush bed and calming colors. The room should incorporate natural materials, such as wood and plants, and feature a limited color palette with soothing hues. The overall atmosphere should be one of calmness and serenity, with a sense of openness and flow between the different areas of the room. The design should prioritize functionality and comfort, creating a peaceful retreat that supports the client’s lifestyle and needs.”

However, even this most satisfactory visualization contained notable inaccuracies: specifically, the room’s original window was mistakenly replaced with a cabinet, and the existing door was unintentionally transformed into a blank wall, leaving the scene with no apparent entrance.

To address these critical spatial errors, I engaged in additional iterative dialogues with LLaMA to clearly define desired window and door designs. I explicitly requested prompts describing a floor-to-ceiling transparent glass window to restore visual openness and a warm, natural wooden door for subtle yet functional access. Utilizing targeted inpainting with the generated prompts, multiple attempts (three for the window and three for the door) were made to restore these architectural elements. Unfortunately, these attempts resulted in

persistent mismatches. The window consistently generated featured heavy curtains instead of the requested clear glass pane, detracting from the intended visual openness.

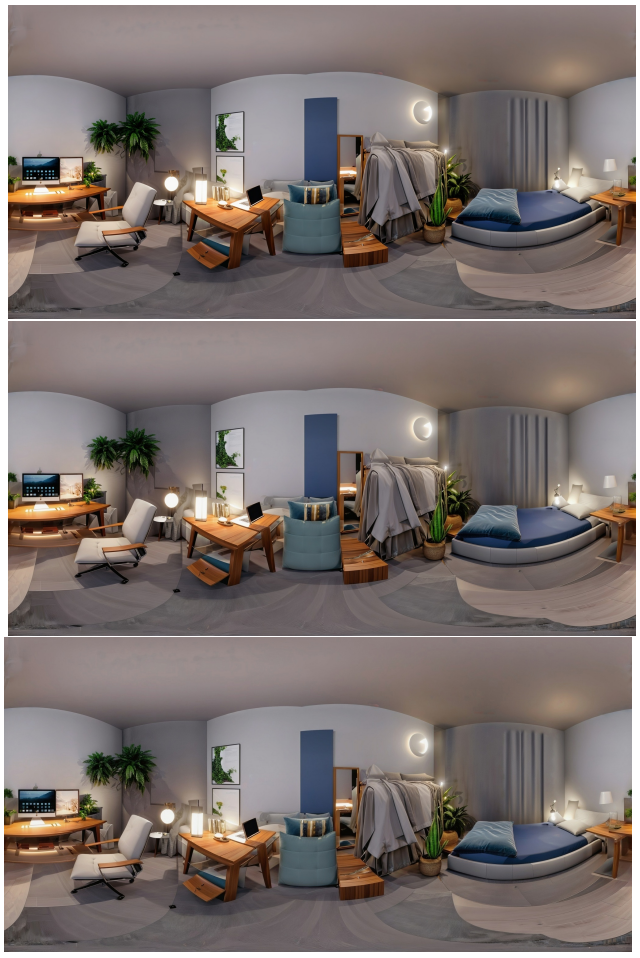


Figure 19: Inpainting results attempting to restore a transparent glass window but resulting in heavy curtains.

Similarly, repeated attempts to introduce an appropriate wooden door via inpainting resulted in abstract forms or indistinct panels, failing to create a coherent and recognizable door structure.



Figure 20: Inpainting results attempting to introduce a wooden door, consistently failing to achieve a clearly defined door form.

Overall, while the initial semantic and thematic exploration proved robust—accurately capturing the client's functional and aesthetic preferences—the experimental outcomes highlighted critical limitations in the system's capacity for targeted spatial modifications and semantic consistency during iterative visual refinement. Despite meaningful suggestions regarding materials, layouts, and lighting, these technical and interpretive shortcomings demonstrated clear areas requiring improvement within MerGen's interactive refinement workflow.

Experiment 2: Minimalist Urban Living **Client Brief:** A young professional couple living in a compact urban apartment seeks to transform their current cluttered space into a minimalist, open-concept living area. Their goal is to maintain a sense of warmth and comfort while maximizing openness and natural light. They prefer neutral tones and natural materials, but do not want the space to feel sterile or underfurnished.

Interaction Summary: The design process began with a dialogue in which I described the client's preferences and goals to the LLaMA agent. The agent responded with stylistic suggestions aligned with minimalism, including the use of white and neutral color palettes, natural wood finishes, and a reduction of decorative clutter.

Next, I uploaded a panoramic image of the current room. The agent correctly identified clutter as a major issue and proposed a layout strategy focused on creating visual openness through restrained furniture selection and negative space.

Following several turns of semantic refinement, I requested a prompt suitable for panoramic image generation. The agent suggested a concise, well-formed prompt emphasizing clean lines, light tones, and natural texture. Based on this, I triggered a full-room scene generation using the Skybox AI engine.



Figure 21: Original panorama image of the room uploaded to the system.

Generation Results and Refinement: The first generated image was immediately promising in both layout and atmosphere, accurately reflecting the minimalist intent. However, I observed that the design, while clean and calm, lacked functional storage. This underscored a typical pitfall of minimalist design—form taking precedence over utility.

I raised this issue in follow-up conversations with the agent, who acknowledged the absence of storage and suggested integrating built-in cabinetry. I then performed three inpainting attempts targeting a specific wall area, asking for a subtle cabinet integration.



Figure 22: Initial generated result based on prompt from LLaMA. Captures the minimalism well but lacks storage functionality.

Despite iterative prompt adjustments and guidance, the inpainting results did not yield a convincing cabinet design. Generated modifications either introduced incoherent elements or altered the spatial logic of the wall. This highlighted a current limitation in precise spatial control using image-to-image generation methods in stylistically constrained contexts like minimalism.



Figure 23: Inpainting attempt to integrate built-in cabinetry into the wall. The result remained visually inconsistent with the minimalist language.

Insights: This experiment showed that the system is capable of generating strong initial layouts that align well with abstract stylistic goals. However, when asked to resolve more nuanced practical concerns—such as integrating invisible storage within a minimalist aesthetic—the system struggled to deliver coherent visual adjustments. This reflects a current trade-off between aesthetic fidelity and spatial utility in AI-generated interior design.

Experiment 3: Designing a Bohemian Living Room from a White-Box Space Client Brief and Spatial Context:

This experiment explores the MerGen system's performance in a "blank canvas" scenario: a white-box interior with only architectural elements (walls, ceiling, floor, windows, and doors), entirely devoid of furniture or decorative elements. The design objective was to create a relaxed, expressive living room in a *Modern Bohemian* style, featuring natural textures, layered soft furnishings, and a cozy atmosphere.

Initial Thematic Exploration:

The design session began with an introductory dialogue describing the spatial void and desired stylistic

direction. The LLaMA agent responded with thoughtful spatial organization ideas, proposing zoning strategies, color palettes, and materials consistent with bohemian aesthetics—such as reclaimed wood, jute, linen, and earthy tones. When asked for a prompt, the agent generated a vivid, richly layered description, emphasizing plush textiles, warm lighting, and relaxed layouts.



Figure 24: Initial panoramic input: a fully empty white-box room.

First Generation and Evaluation:

Using the prompt, a panoramic visualization was generated through Skybox AI. The result captured the desired ambiance and styling accurately, showcasing relaxed furniture arrangements, rich textures, and layered lighting. This demonstrated the system's strength in extrapolating thematic concepts from abstract briefs even without existing furniture cues.



Figure 25: Initial generation result using LLaMA-generated prompt for a modern bohemian interior.

Iterative Enhancement:

Despite stylistic fidelity, the scene appeared somewhat sparse due to the initial white-box input. Through further conversation, the agent suggested adding a bohemian-style sofa. The user then requested an inpainting prompt and performed targeted mask editing to modify a wall section with a new sofa. The inpainting result preserved visual coherence and successfully enriched the room's composition while maintaining the overall bohemian theme.



Figure 26: Mask-based inpainting result to add a bohemian-style sofa.

Reflections:
This experiment confirmed MerGen's effectiveness in early-stage sketching for a fully undefined interior, showcasing the system's capability to initiate design from scratch and deliver coherent stylistic outcomes. The integration of conversational prompting with visual generation and spatial modification proved especially valuable when no pre-existing context is available.

Experiment 4: Transforming a Living Room into a Social Interaction Hub **Client Profile:** Two roommates sharing an urban apartment. **Design Objective:** Transform a typical living room into a flexible, sociable environment that supports informal gatherings, collaborative work, and small group dining. **Design Keywords:** informal, vibrant, movable seating, warm lighting, multifunctionality.

This experiment specifically investigated whether the MerGen system could support spatial design goals that center on functional transformation—namely, reprogramming a space to enable a new social use. The initial panoramic input depicted a traditional living room centered around a sofa and TV, with a layout unsuitable for hosting or dynamic interaction. **Figures:**



Figure 27: Original input: Conventional living room panorama with sofa-TV setup.

I began by discussing the clients' social needs with the LLaMA agent, including the desire to seat five or six people comfortably, allow flexible arrangements, and maintain a lively yet relaxed mood. The agent responded with appropriate spatial ideas, such as introducing bean bags, portable tables, accent lighting, and adding visual anchors like rugs or color blocks to promote social zoning.

However, despite these rich conversational exchanges, the panoramic scene generated from the agent's prompt failed to reflect any of the discussed functional goals. While the lighting and palette aligned with the

mood, the furniture layout remained rigid and solitary in feel. The image resembled a private lounge rather than a space for group activity.



Figure 28: Four Generated panoramic scene reflecting the intended mood, but all lacking social layout transformations.

To address this mismatch, I asked the agent to analyze the generated result. The agent again recommended introducing low, moveable seating and lounge-style furnishings, such as a large bean bag in a muted color. I attempted to apply this advice using targeted inpainting on a clear zone of the room. Nevertheless, three separate attempts to generate the bean bag resulted in either vague surface noise or no noticeable addition.



Figure 29: Inpainting attempt to add a bean bag. No clear result achieved, despite targeted prompt.

This outcome highlights one of the current limitations in MerGen's workflow: while the dialogue interface is capable of handling abstract programmatic goals and offering appropriate interventions, the connection between functional design logic and visual representation remains inconsistent. Despite promising semantic alignment during interaction, the system lacked sufficient control and grounding to successfully translate these ideas into spatial form.

Experiment 5: Freshening Up an Aging Flower Shop Material Selection and Specification

In this experiment, I explored how MerGen supports material and visual enhancement in a small commercial setting—an aging flower shop seeking a low-budget, fresh, and natural look. Following the initial full-room generation, which successfully brightened the space and improved its visual appeal, the AI agent was asked to suggest strategies for making the entrance more inviting.



Figure 30: Original flower shop image before transformation.



Figure 31: Initial Skybox generation reflecting a fresher, more natural shop atmosphere.

The LLaMA agent proposed improvements to the entrance area, emphasizing the use of natural decorative elements. Upon receiving the suggestion to place a potted plant near the door, I used mask-based inpainting with the provided prompt. Although the final result did not substantially alter the sense of invitation, it did reflect the agent's suggestion by placing a small planter near the entryway.



Figure 32: Inpainting result after entrance improvement, including addition of a potted plant.

Subsequent conversations focused on material refinement. While I did not explicitly request detailed wallpaper suggestions in this experiment, the visual update of the scene implicitly included recommendations aligned with a natural, handmade aesthetic—such as muted wall colors and wooden textures—aimed at maintaining the "human warmth" of a long-established shop while improving clarity and freshness.

The overall outcome demonstrated MerGen's capacity to recommend atmosphere-enhancing elements and simulate soft decoration changes through inpainting. However, limitations remained in achieving more abstract goals such as "inviting feel" or "emotional warmth," which were not always effectively conveyed through generated visuals alone.

4.3.4 Key Findings from Experiments

Across all five scenarios, the following key findings emerged:

- **Creative Guidance:** The LLaMA agent consistently provided meaningful stylistic suggestions that expanded the design space while remaining faithful to the user's initial intent. Its responses reflected domain-specific language and professional reasoning, reinforcing its role as a virtual design partner.

- **Semantic Clarity:** Dialogue-driven interactions were successful in translating high-level design concepts into thematically coherent and spatially plausible panoramic images. Prompts generated through iterative dialogue accurately conveyed visual intent.
- **Effective Iteration:** The MerGen workflow enabled fast transitions between conversation, scene generation, and targeted inpainting. This allowed for efficient testing of multiple design alternatives and offered a novel form of ideation-through-dialogue.
- **Design Agency and Control:** Despite leveraging AI for guidance, the process maintained strong user autonomy. Designers retained control over prompt formulation and modification targets, ensuring AI remained a supportive rather than prescriptive agent.
- **VR-Enabled Decision Support:** Immersive VR visualization played a critical role in supporting rapid judgment and decision-making. By viewing AI-generated outcomes as full panoramic scenes within a spatial context, users could immediately assess layout quality, visual comfort, and stylistic consistency. This approach brought abstract design thoughts into a realistic and embodied simulation, allowing decisions to be grounded in an intuitive, human-scale representation of space.

However, the experiments also revealed key limitations. The agent's image analysis remained general rather than critical; it frequently confirmed generated outputs as "aligned" even when key elements were missing or altered (e.g., windows replaced by cabinets). Attempts at structural inpainting often failed, with the system misinterpreting spatial intent (e.g., inpainting windows resulting in curtains). While stylistic language and overall mood alignment were strong, precise spatial manipulation remained a challenge. These observations reinforce the need for clearer feedback mechanisms and improved understanding of spatial-functional logic in future systems.

Experiments	Theme & Goal	Key Actions	Main Issues	Key Findings	Generations
1. Serene Bedroom Transformation	Design a calm bedroom-office hybrid	- Dialogue-driven zoning discussion - Panoramic visualization - Initial scene analysis	- Missing window and door - Misinterpretation of some terms and objects	**Successfully captured atmosphere**; **struggled with inpainting** to correct partial generation.	
2. Minimalist Urban Living	Create a minimalist, open-plan urban apartment	- Decluttering strategy - Prompt refinement - Panoramic generation	- No built-in storage in generated image - Difficulty adding subtle cabinets	Strong minimalist **styling achieved** **Lacking functional storage solutions. Inpainting results are not ideal**	
3. Modern Bohemian Living Room	Build a cozy, layered bohemian space from a white-box	- Style exploration - Panoramic generation - Sofa addition via inpainting	- Sparse initial layout - Need for richer furnishing	**Effective theme translation from scratch**, **Successful inpainting generation of object.**	
4. Social Interaction Hub	Redesign a living room for dynamic social use	- Multi-user flexibility goals - Space analysis - Panoramic generation	- Static, isolated layouts - Failed to introduce movable seating	**Good analysis and suggestion** **Functional transformation not expressed in generation**	
5. Flower Shop Refresh	Refresh a flower shop with airy, fresh aesthetics	- Style briefing - Panoramic generation - Entrance adjustments with inpainting	- Minor distortions inpainting refinement	**Overall thematic transformation achieved**; small object **inpainting not very ideal**.	

Figure 33: Experiment results of module 1.

4.3.5 Addressing AIA survey Concerns

The experimental results also reflect how the MerGen system begins to address major adoption concerns identified in the AIA report regarding AI in architectural practice:

- **Reliability:** In most cases, AI-generated outputs maintained strong alignment with clearly stated design intentions, particularly during initial sketching. However, breakdowns occurred during fine-grained inpainting, where control fidelity was limited. This highlights areas for further development in spatial precision and prompt grounding.
- **Transparency and Interpretability:** LLaMA's ability to provide reasoning through dialogue offered a novel form of design transparency. Users could inquire about design choices, request clarifications, or steer conversation direction, resulting in improved understanding of why certain stylistic decisions were made. Moreover, the immediate visual feedback from panoramic generation and inpainting enhanced interpretability.
- **Iterative Flexibility and Creative Support:** The conversational modality, combined with visual generation and gesture-based inpainting, created an open-ended iterative environment. Designers were empowered to test ideas quickly, receive AI-suggested alternatives, and refine their vision in a fluid loop. This structure aligns closely with how early-stage ideation unfolds in professional practice.
- **Embodied Evaluation through VR:** One of the clearest strengths of the MerGen system was its immersive deployment. Unlike traditional flat-screen representations, the VR context enabled real-time judgment of spatial relationships, lighting, scale, and compositional balance from a first-person viewpoint. This embodied perspective allowed users to interpret AI suggestions not just as images, but as lived spatial experiences, thus reducing abstraction and enhancing decision confidence.

While the prototype does not yet achieve full parity with human design partners, it makes significant strides toward augmenting conceptual workflows. It provides an approachable entry point for non-AI experts, enhances thematic exploration, and lays a foundation for future systems that could operate as collaborative, transparent design aides.

4.4 Material Selection and Specification

4.4.1 Research Goals and Rationale

Material selection plays a critical role in architectural design, directly shaping the spatial atmosphere, functional performance, and aesthetic coherence of an environment. Unlike early sketching, which prioritizes conceptual ideation, this phase of the evaluation focuses on grounded decision-making—translating abstract spatial narratives into tangible design elements such as wall finishes, furnishings, and surface treatments [2].

The objective of this stage was to assess the MerGen prototype's ability to support detailed material exploration and specification. This included three key capabilities: (1) providing stylistically aligned and contextually relevant material suggestions through natural language dialogue with the LLaMA agent [15]; (2) leveraging Retrieval-Augmented Generation (RAG) to offer concrete examples of real-world materials and

products [30]; and (3) using targeted inpainting to visualize material substitutions and refinements directly within the panoramic scene [35].

This phase also tested the system's ability to maintain semantic continuity across a multi-turn session, ensuring that material choices aligned with earlier design intentions established during the sketching phase. For instance, the agent was expected to recommend wallpaper textures, furniture types, or surface colors that matched previously discussed themes such as minimalism, warmth, or artistic expression [28]. In addition, the system was tasked with summarizing material decisions into clear, specification-style outputs—simulating a common step in real-world architectural workflows where designers compile annotated summaries of product recommendations, finishes, and pricing [19].

Overall, this evaluation sought to determine whether MerGen could meaningfully bridge the gap between conceptual design and implementation-level detailing, and to what extent its AI-driven recommendation pipeline was interpretable, controllable, and useful within an immersive VR-based workflow.

4.4.2 Methodology and Workflow

The material specification phase of each experiment followed a consistent, multi-stage process designed to simulate a realistic interior design workflow. This phase was not independent but served as a direct continuation of the early-stage design exploration described in Section 4.3. Each of the five experiments in this section corresponds one-to-one with the scenarios introduced earlier—carrying forward the same spatial context, design objectives, and stylistic direction—while shifting focus to material articulation, furniture selection, and final specification. The methodology comprised the following steps:

1. **Conversational Material Exploration:** I initiated a focused dialogue with the LLaMA agent to discuss material options—such as wallpapers, flooring, or furniture finishes—based on the previously generated panoramic scene. The goal of this dialogue was to refine the material palette in relation to style, mood, functionality, and user-defined constraints (e.g., budget, atmosphere, or maintenance needs).

2. **Retrieval of Real-World Examples (RAG):** Once a general material direction was established, the agent attempted to retrieve concrete examples of relevant products using a Retrieval-Augmented Generation (RAG) pipeline. This process drew from a curated knowledge base of product descriptions (e.g., IKEA catalog entries), aiming to surface image-based examples complete with pricing and summary descriptions. This step was intended to simulate material sourcing as it occurs in real-world practice.

3. **Visual Application via Inpainting:** To test how suggested materials would appear in context, I used Unity's VR interface to mark areas of interest (e.g., a wall, table, or floor) using a mask painting tool. These masks, along with a refined prompt, were submitted to the inpainting module, which visually replaced the masked areas in the panoramic image with AI-generated alternatives. This step enabled immersive visual verification of proposed changes without needing full-scene regeneration.

4. **Specification Summary Generation:** After material options were discussed and tested, I instructed the LLaMA agent to generate a formal design specification summary. This included style rationale, rec-

ommended materials, estimated pricing, and brief justifications—mirroring the kind of documentation typically prepared for clients or project teams during the later stages of interior design.

This workflow allowed for iterative and semantically consistent transitions from abstract dialogue to visual decision-making and formal documentation, bridging the gap between conceptual ideation and practical specification within an immersive design context.

4.4.3 Detailed Experimental Narratives

Experiment 1: Serene Bedroom Transformation In the second phase of the experiment, following the completion of the spatial and stylistic layout, I initiated a dialogue with the LLaMA agent to explore appropriate wall finish options for the serene multifunctional bedroom. The primary goal was to assess whether the system could offer not only stylistically relevant recommendations but also retrieve real-world material examples to support decision-making.

The agent provided several high-level suggestions for wallpaper materials. These included recommendations such as muted grasscloth textures, natural fiber finishes, soft geometric patterns, and non-woven vinyl alternatives. These suggestions aligned conceptually with the previously established design goals—specifically the emphasis on texture, softness, and tonal restraint.

> *"A soft, muted grasscloth wallpaper in a neutral tone such as beige or light gray would bring warmth and texture to a room... A textured, non-woven wallpaper in a monochromatic color scheme would add depth and visual interest."*

However, when prompted to retrieve specific examples using the retrieval-augmented generation (RAG) module, the system consistently failed to surface appropriate products. Instead of wallpaper, each of the three retrieval attempts returned unrelated furniture items, most notably a wooden table, demonstrating a disconnect between the semantic intent of the prompt and the structure of the underlying vector database. This indicated significant limitations in the current RAG implementation, particularly in its ability to semantically parse and filter design-specific categories such as surface finishes.



Figure 34: When asked to retrieve wallpaper product, the agent returned a table item.

While the agent's narrative descriptions were stylistically appropriate and demonstrated sensitivity to the overall design language, they remained general in nature and failed to translate into actionable specifications. When asked to summarize the full session as a structured specification document, the LLaMA agent only included the wallpaper-related decisions and omitted earlier discussions around layout zoning, window treatment issues, and prompt-based spatial generation. This omission further revealed a limitation in the system's memory summarization and session tracking mechanisms.

Taken together, this phase of the experiment highlighted two key challenges: first, that general material suggestions may align in tone but lack actionable specificity; and second, that retrieval-based augmentation requires stricter control over taxonomy and embedding precision to produce reliable outputs. Despite a conceptually aligned design intent, the inability to retrieve, filter, and reason about product-level data constrained the prototype's usefulness in material selection workflows.

Experiment 2: Minimalist Urban Living In the material selection phase of this experiment, I initiated a follow-up dialogue with the LLaMA agent to explore appropriate wall treatments for the minimalist living room generated in the previous design stage. The goal was to determine whether the system could effectively recommend and help visualize materials that matched the established design language—namely, visual simplicity, spatial clarity, and a neutral color palette.

Upon asking "What kind of wallpaper would you suggest for this room?", the agent offered a reasonable and stylistically aligned response: recommending subtle geometric patterns, natural fiber looks (such as grasscloth or linen), and neutral tones like white, gray, or beige. These suggestions aligned well with the client's minimalist intent and offered helpful guidance for textural variation without introducing visual clutter.

Following this, I triggered the system's Retrieval-Augmented Generation (RAG) mechanism to retrieve real-world product suggestions based on the descriptive output. However, the system repeatedly returned

irrelevant results—in this case, a pine wood table rather than wallpaper. This mismatch between prompt and retrieval result revealed a notable gap in MerGen's RAG configuration and indexing quality for material types. The agent was unable to provide a usable wallpaper product with corresponding image, price, and description, thereby limiting its value for real-world material sourcing in this scenario.



Figure 35: Same as experiment 1, When asked to retrieve wallpaper product, the agent returned the same table item.

To evaluate the agent's ability to document design decisions, I requested a formal summary and a "SPAG book" style output. While the agent was able to produce a clear summary listing the room's main features—including the aesthetic rationale, suggested cabinetry, and wallpaper characteristics—the output remained relatively high-level. It lacked specificity regarding product dimensions, brands, or installation notes typically found in architectural material specifications.

Overall, this phase demonstrated the agent's strengths in stylistic reasoning and abstract material language, but also exposed limitations in product-level specificity and semantic-to-product mapping. While the dialogue flow supported conceptual decision-making, the inability to generate accurate or diverse material examples through the RAG system hindered its potential as a practical design assistant in this context.

Experiment 3: Bohemian Living Room from a White-Box Space In this experiment, I evaluated MerGen's material recommendation capabilities within the context of a Modern Bohemian living room designed from an empty white-box environment. The aim was to verify whether the system could provide thematically appropriate suggestions and realistic examples of furnishings and finishes, especially when starting from a blank spatial slate.

After the initial design generation and inpainting of a sofa element, I initiated a focused material dialogue with the LLaMA agent by asking for wallpaper suggestions to match the bohemian theme. The agent

recommended a natural, earthy wallpaper with subtle texture—such as linen or grasscloth styles—highlighting its compatibility with organic, relaxed design aesthetics.

To test the retrieval capability, I triggered the system's RAG module to return relevant product examples. This time, the system correctly retrieved a product classified as wallpaper, with a suitable description and pricing. This outcome represented a notable improvement in semantic alignment between material intent and retrieval content, contrasting earlier experiments where unrelated items were returned.



Figure 36: When asked to retrieve wallpaper product, the agent returned a wallpaper product image and its price of "$6.60 / sq ft".

Following this, I asked the agent to generate a final specification summary based on the entire design process. The output successfully consolidated key elements: a linen-upholstered sectional sofa, jute rug, reclaimed wood coffee table, natural-textured wallpaper, and ambient lighting via string lights and linen-shaded lamps. It conveyed a coherent understanding of the overall style and design rationale.

However, the agent's attempt to estimate pricing across all elements lacked grounding in actual retrieved product data. Cost ranges for furniture and décor items were presented ($500–$2,000 for the sofa, $100–$500 for wallpaper and rug), but without justification or citation from RAG of any earlier discussion, which might be a sign of hallucination. This highlighted a current limitation in price traceability and transparency within MerGen's material specification capabilities.

Overall, this experiment demonstrated the system's potential for end-to-end material exploration and design documentation, particularly in translating abstract stylistic goals into tangible material suggestions. The specification output showed clear thematic alignment, though future versions could benefit from tighter integration between retrieval results and pricing documentation.

Experiment 4 (Social Interaction Living Room) In the material specification phase of this experiment, I engaged the LLaMA agent in detailed discussions about specific furnishing and decor elements, focusing on how to reinforce the intended social functionality of the living room.

The conversation began with suggestions about wallpaper. The agent proposed several directions, including botanical prints, textured neutrals, and bold geometric patterns, all intended to balance the lively mood of the space while avoiding visual overload. When asked to retrieve examples using the RAG system, it returned three wallpaper option with price and description. These option was semantically aligned with earlier discussions and visually appropriate, though not grounded in actual vendor data.



Figure 37: When asked to retrieve wallpaper product, the agent returned a same wallpaper product image as in experiment 3 and its price of "$6.60 / sq ft".

I also requested recommendations for a coffee table. The agent advised using warm-toned, natural materials—such as reclaimed wood or live edge designs—to support the earthy aesthetic. The retrieval system successfully returned a suitable example: a plank-constructed coffee table priced at $799, reinforcing the agent's design direction and matching the previously generated scene.



Figure 38: Same as experiment 1, When asked to retrieve coffee table product, the agent returned a wood table item as discussed.

Although the RAG-based retrieval yielded more relevant items in this experiment compared to earlier ones, the final specification summary remained overly abstract. It described the living room as featuring a "modular seating arrangement," "earthy tones," and "natural and synthetic materials," yet omitted specifics about individual selections or their spatial relationships. Furthermore, the cost breakdown was limited to general ranges and lacked itemized pricing or sourcing.

This experiment demonstrates that while the MerGen system can increasingly provide contextually relevant product suggestions and refine them through conversation, it still struggles to compile a truly comprehensive, actionable specification. The gap between semantic discussion and structured, detailed documentation remains an area for further enhancement.

Experiment 5: Freshening Up an Aging Flower Shop Following the generation of a refreshed panoramic image for the flower shop in the last module, I proceeded to refine the material specifications. The LLaMA agent suggested applying a natural, earthy wallpaper with a subtle botanical pattern to reinforce the welcoming and organic aesthetic. The stylistic advice emphasized hand-drawn botanical motifs in muted tones such as sage green and sandy beige, aiming to preserve the store's warmth while modernizing its visual identity.

When prompted for concrete product examples, the system successfully returned a matching wallpaper sample using the RAG module. The retrieved example featured appropriate description and pricing, and visually aligned with the recommended material qualities.



Figure 39: Wallpaper suggestion retrieved through RAG: A poetic design that brings a sense of calm into your space with a delicate botanical pattern that feels timeless and romantic.

To integrate this design recommendation, I used inpainting to alter the existing wall surface. However, the modification failed to convincingly represent the intended wallpaper. Instead, the AI introduced decorative elements—such as hanging greenery—that did not constitute a material replacement. While visually pleasant, this result pointed to current limitations in using inpainting for precise surface transformations.



Figure 40: Inpainting result intended to apply wallpaper. The modification added green accents but failed to change surface material.

In the final step, I requested a design specification summary. The output provided a high-level overview of the design direction—highlighting natural materials, warm lighting, and a botanical ambiance—but the description of chosen products remained vague and lacked quantitative support for pricing or sourcing. This

reflects a broader challenge in AI-generated summaries: maintaining coherence while integrating specific decision history.

4.4.4 Key Findings from Experiments

The material selection and specification phase across all five design scenarios revealed both the strengths and current limitations of AI-assisted workflows for architectural specification. The following key insights were observed:

- **Material Reasoning and Stylistic Suggestions:** The LLaMA agent generally provided thematically appropriate material suggestions based on spatial context and stylistic goals. It was capable of offering abstract advice (e.g., "grasscloth wallpaper" or "reclaimed wood table") that aligned well with the scenario's intended atmosphere and tone.
- **Retrieval of Product Examples via RAG:** The retrieval module successfully returned concrete items in most cases, including images, descriptions, and estimated prices. However, retrieval quality remained inconsistent: in some cases, the returned items mismatched the query (e.g., returning a wooden table when asked for wallpaper), suggesting limitations in the embedding space or vector search precision.
- **Visual Feedback through Inpainting:** The inpainting system showed limited effectiveness for material changes. While it occasionally succeeded in inserting decorative items (e.g., curtains, plants), it struggled to modify wall textures or surface finishes accurately. Instead of changing material appearance, it often added unrelated visual elements or failed to localize the target area precisely, making it unsuitable for controlled material application.
- **Specification Generation:** The agent could be prompted to generate summaries of prior design decisions, but the outputs varied widely in quality. Some responses were overly vague or lacked critical details, while others introduced hallucinated content not present in prior discussions. This inconsistency undermined the reliability of the specification output as a design documentation tool.
- **VR as an Evaluation Medium:** The immersive VR environment added substantial value to the evaluation process. Being able to view AI-generated materials or modifications within a spatially accurate panoramic context helped inform judgment about material suitability and visual coherence. Immediate visual feedback from standing inside the design space enhanced decision-making compared to traditional flat-screen workflows.

While the prototype supported exploratory material reasoning and helped visualize possibilities within immersive scenes, it currently lacks the fidelity and control necessary for precise specification or downstream documentation workflows

Experiments	Focus in Material Stage	Material Discussion with AI	Product Retrieval and Inpainting	Spec Book Output & Issues
1. Serene Bedroom Transformation	Emphasis on soft natural textures, calm palettes	- Discussed materials like wood, woven fabrics - Lighting recommendations	Rag **failed** to return a wallpaper product when asked	Spec generation covered zones and atmosphere, but **lacked detailed material breakdown** **Included products that didn't exist in the conversation**.
2. Minimalist Urban Living	Balancing minimalism with hidden storage	- Explored built-in storage solutions - Neutral tones and natural finishes	Rag **failed** to return a wallpaper product when asked	Spec information **lacks details of reasonings**.
3. Modern Bohemian Living Room	Layering rich natural materials	- Discussed layered fabrics, warm woods - Suggested accent pieces	Rag **successfully returned** a wallpaper product when asked	Spec **successfully captured atmosphere, furniture types, and rough material types**.
4. Social Interaction Hub	Creating flexible, sociable furniture layouts	- Suggested modular seating, lightweight movable items - Emphasis on vibrant, informal tone	Rag **successfully returned** a wallpaper product when asked Rag **successfully returned** a coffee table when asked	**Spec document described desired flexibility** **Included products that didn't exist in the conversation**.
5. Flower Shop Refresh	Light material palette emphasizing plants	- Recommended lightweight shelves, open displays - Natural light emphasis	- Rag **successfully returned** a wallpaper product when asked - **Failed on inpainting for material replacement**	Spec **effectively summarized** mood, lightness, and material recommendations for renovation.

Figure 41: Experiment results of module 2.

4.4.5 Addressing AIA survey Concerns

This phase of the evaluation addresses several key concerns raised in the AIA report regarding the integration of AI into professional architectural workflows, specifically in the context of material decision-making and design documentation:

- **Reliability of Recommendations:** The AI system demonstrated an ability to generate thematically consistent and stylistically appropriate material suggestions. However, in several cases, especially when using RAG for product retrieval, the outputs lacked specificity or returned unrelated items. This indicates that while semantic understanding is promising, the retrieval backend still requires improvements to be dependable in real-world specification contexts.

- **Transparency and Interpretability:** Through natural language dialogue, the AI provided rationales for material choices, often referencing texture, color, and spatial harmony. These conversational explanations enhanced interpretability for the designer. Yet, the lack of detailed sourcing or verifiable references for product data limited full transparency, especially for cost estimations or brand-specific recommendations.

- **Iterative Visual Feedback:** The prototype supported real-time visual testing through inpainting, intended to preview material changes in-situ. However, the inpainting model showed poor performance on material substitution tasks, such as wallpaper or surface texture updates. Its strength appeared more in object-level edits rather than material-layer alterations, which reduced its utility for precise feedback and undermined visual verification of material changes.

- **Specification Support and Documentation:** While the AI could generate summaries of material decisions, the results were often too generalized, occasionally inaccurate, and not reliably structured for professional use. Some outputs included hallucinated elements or failed to include key decisions made earlier in the dialogue, highlighting a major gap between conversational summarization and actionable specification output.
- **Immersive Validation through VR:** Despite these limitations, the VR interface played a significant role in bridging the gap between abstract material discussions and experiential understanding. The immersive visualization allowed for immediate spatial and aesthetic feedback, supporting faster iteration and clearer decision-making compared to traditional selection interfaces. It provided a practical layer of validation for otherwise abstract AI recommendations.

In summary, while the MerGen system shows potential in addressing the AIA survey's concerns about transparency, and provides creative augmentation, and immersive visualization, it currently falls short in delivering the technical reliability and precision required for material documentation and professional-level specification. These limitations mark clear directions for future improvement.

5 Results

5.1 Preliminary Findings

This research investigated how the convergence of generative AI and VR technologies can support architectural design activities in early-stage ideation and material specification. The results across five design scenarios reveal several meaningful outcomes relevant to real-world design practices and directly respond to the guiding research questions.

Research Focus	**Module Addressing It**	**Summary**
Real-time Interactivity (SubQ1)	Module 1	AI dialogue + VR visualization helped early-stage design with real-time interaction. ● Dialogue with AI agent helped produce ideas and generated panoramic images that align with design intentions. ● VR environment allowed immersive, full-space visualization of ideas, and helped with decision making. ● Iterative refinement enabled continuous design improvements.
Technical and Workflow Challenges (SubQ2)	Module 1&2	● Sometimes there are misinterpretations of user's speech; ● Inpainting-based Refinements are not working ideally; ● Generation Delays decreased the fluency of workflow; ● RAG Limited (database size, semantic retrieval mismatches)
Actionable Deliverables (SubQ3)	Module 2	RAG precision and hallucination lead to unstable specification quality, sometimes missed or inaccurately described material and product details.
Overall Pipeline impact (Main RQ)	Module 1&2	Demonstrated a continuous pipeline from conceptual ideation to specification.

Figure 42: Reflections on research questions.

In terms of early-stage sketching and thematic exploration, the integration of conversational AI and panoramic visualization allowed users to externalize and refine abstract design concepts through a structured, dialogue-driven workflow. In all five experiments, semantic conversations led to generation prompts that reflected key design intentions such as minimalism, social interaction, or emotional atmosphere (e.g., freshness, calmness). In most cases, the initial image generations effectively captured the specified style and spatial language. This demonstrates that AI, when grounded by contextual dialogue, can support designers in quickly translating vague ideas into immersive visualizations.

The immersive nature of the VR environment played a critical role in shaping user decision-making. Working inside the generated scene enabled spatial assessments that would be difficult through conventional 2D sketches. Users could directly evaluate visual proportions, perceived openness, and lighting qualities. This visual feedback loop—enabled by immersive embodiment—led to faster identification of undesired features and clearer articulation of subsequent changes. The platform thus enabled not only quicker iteration cycles, but also more confident spatial reasoning grounded in realistic representations.

Material specification and refinement was supported through a combination of AI-suggested semantic reasoning and RAG-based retrieval from a curated database. The conversational agent provided relevant

suggestions regarding material type, tone, and texture, which were then followed by attempts to retrieve real-world product examples (e.g., wallpapers, furniture). In several cases, the retrieved items aligned well with the described goals. However, when tested further through inpainting-based visual application, the results were more mixed. While AI was generally effective at inserting new objects (e.g., plants, sofas), its ability to transform surfaces or materials was limited, and in some experiments, editing attempts failed to render the intended changes.

Specification generation, meant to bridge conceptual design and documentation, was also tested through natural language prompts. While AI could produce summaries of style and selected materials, the generated specifications were often too abstract, occasionally omitting discussed decisions or fabricating plausible-sounding but unverified pricing data. This highlights that while AI can assist in documenting design thinking, further structure and verification layers are needed to produce truly actionable deliverables.

Overall, the results affirm that AI can serve as a valuable partner in early conceptual phases of design—especially when guided by clear semantic intent and combined with immersive spatial evaluation. However, downstream tasks like material verification, specification generation, and detailed editing remain areas of partial success, limited by current model capabilities and dataset coverage.

5.2 Limitations

While the results validate the potential of AI and VR integration in small-scale architectural design, several limitations were observed that constrain the current framework's scalability and reliability.

One key limitation lies in the inpainting tool used for scene editing. While relatively successful in inserting discrete objects such as furniture or lighting, it struggled to perform fine-grained material transformations. For example, attempts to alter wall finishes or add new architectural details via inpainting often resulted in irrelevant visual artifacts or inconsistent textures. This highlights the limitation of using image-based tools for tasks that may require semantic understanding of spatial boundaries or structural hierarchy.

The Retrieval-Augmented Generation (RAG) system used for product suggestion was constrained by dataset size. The current database consisted of approximately 20 examples each for wallpaper and table products. While useful for demonstrating proof of concept, this narrow scope limited the variety and specificity of recommendations. In some cases, unrelated items were returned (e.g., a pine table for a wallpaper request), indicating weak semantic alignment between input descriptions and indexed content. Expanding and curating this database with richer metadata will be critical for enabling robust material recommendations.

Conversational misrecognitions from speech-to-text transcription also introduced workflow interruptions. In several experiments, key phrases were misinterpreted (e.g., "the client also wants" transcribed as "decline also one"), leading to irrelevant responses. These errors complicate dialogue flow and undermine the sense of seamless interaction. Customized voice recognition models trained on design-specific vocabulary may be necessary for future refinement.

Finally, although the platform successfully generated scene summaries and specifications on request, these outputs lacked precision. Many included only broad generalizations, omitted discussed features, or proposed imaginary price ranges. This indicates a need for structured prompt engineering and stronger connections

to verified product databases or material libraries if AI is to be trusted for deliverables like SPAG books or client-facing documents.

Despite these limitations, the system demonstrated tangible benefits in early design workflows and spatial visualization. Future development must address the transition from inspiration and ideation to accurate, verifiable, and editable outputs that can support end-to-end architectural workflows.

6 Discussion and Conclusion

6.1 Contributions

This thesis investigates how the integration of generative AI and virtual reality can reshape the architectural design process, with a specific focus on small-scale spatial projects. Through a series of experimental design tasks, a custom AI-VR platform was tested to assess its ability to support ideation, semantic guidance, visual refinement, and material-level specification. The contributions of this research can be summarized across several dimensions aligned with the guiding research questions.

First, the work demonstrates that generative AI can effectively enhance real-time interactivity during design ideation. Through conversational engagement with an interior-design-specialized AI agent, users were able to explore abstract spatial ideas, iterate on design intentions, and translate these ideas into image generation prompts. The dialogues were not merely transactional but showed the agent's ability to suggest stylistic alternatives, respond to critique, and maintain coherence with the evolving design context. This form of semantic feedback loop—where design concepts evolve through verbal reasoning and visual confirmation—marks a significant shift from static generation tools to truly interactive design partners.

Second, the research identifies and partially addresses technical and workflow challenges in integrating AI components within a VR-based design pipeline. A modular backend was developed to unify voice interaction, semantic processing, panorama image generation, inpainting-based edits, and retrieval-augmented product suggestion. The implementation reveals practical integration strategies, such as containerized service orchestration and cross-modality data handling between image and text. At the same time, the work exposes limits in current tool capabilities—such as high latency in scene generation, insufficient mask precision in inpainting, and difficulty maintaining material or object consistency across editing cycles.

Third, the system demonstrates how immersive AI interactions can begin to produce actionable design outputs. Within each experiment, the AI agent was prompted to generate natural language summaries of the design process, including furniture and material recommendations, stylistic rationales, and cost estimates. Although current outputs remain incomplete or overly generalized, the experiments show that such summaries can serve as first drafts of specification documents—bridging early ideation with downstream decision-making. This illustrates the broader potential for AI-VR systems not only to inspire but also to document and communicate design thinking in real-time.

Finally, the research validates that such a system can function across multiple design domains. The five conducted experiments—ranging from minimalist living rooms to flower shop revitalization—demonstrate the transferability of the workflow to diverse client needs and spatial typologies. In each case, the AI was able to maintain a stylistic voice, provide context-aware design suggestions, and facilitate iteration, all while preserving the immersive advantages of working within a 360° spatial representation of the environment.

6.2 Next Steps

While the prototype and experiments offer compelling insights into the role of AI in immersive design workflows, they also reveal opportunities for further development and refinement.

First, to improve the responsiveness and relevance of generative AI suggestions during early-stage ideation, future work should focus on improving the precision of language understanding and visual generation. In several experiments, misinterpretations of spoken input or hallucinated image outputs led to misalignment with the user's design goals. Custom-tuned speech models and tighter grounding between visual generation models and semantic prompts would improve the consistency and reliability of the AI assistant.

Second, the material recommendation component—powered by Retrieval-Augmented Generation (RAG)—is currently constrained by a very limited dataset (e.g., only 20 wallpaper and 20 table items). Expanding the scale, diversity, and metadata richness of this dataset would allow the AI to provide more realistic, diverse, and situation-specific recommendations. This is especially crucial in material-sensitive domains like interior design, where texture, pattern, price, and installation practicality are all important.

Third, inpainting as a form of visual editing must be rethought for architecture-oriented workflows. While effective for inserting new objects such as furniture, inpainting was often ineffective for material transformation (e.g., modifying walls or flooring). Future work might explore geometry-aware editing, segmentation-based control masks, or even hybrid techniques combining parametric modeling with generative fills.

Fourth, the generation of deliverables—such as specifications and SPAG (Style, Product, Application, Guidance) summaries—must evolve toward more structured and complete outputs. Current results often omit key decisions or fabricate details. Future iterations should support schema-driven generation of design documentation, possibly integrating with building information models (BIM) or CAD-based workflows for greater fidelity and traceability.

Lastly, the platform's applicability to larger teams and diverse design contexts remains an open direction. The current framework is tested in a single-user setting. Extending the platform to support collaborative design, multi-agent negotiation, or real-time feedback from external stakeholders (e.g., clients, consultants) could significantly increase its utility in professional practice.

Taken together, these next steps will help transition the system from a promising prototype to a robust framework capable of supporting real-world architectural workflows. More importantly, they will help clarify the evolving relationship between artificial intelligence, spatial creativity, and immersive human-computer interaction in the built environment.

References

[1] AGIRACHMAN, F. A., SHINOZAKI, M., KOERNIAWAN, M. D., AND INDRAPRASTHA, A. Implementing affordance-based design review method using virtual reality in architectural design studio. *Buildings 12*, 9 (2022), 1296.

[2] ALATTA, R. A., AND FREEWAN, A. Investigating the effect of employing immersive virtual environment on enhancing spatial perception within design process. *ArchNet-IJAR: International Journal of Architectural Research 11*, 2 (2017), 219.

[3] AMER, N. A. Architectural design in the light of ai concepts and applications. *MSA Engineering Journal 2*, 2 (2023), 628–646.

[4] AYDIN, S., AND AKTAŞ, B. Developing an integrated vr infrastructure in architectural design education. *Frontiers in Robotics and AI 7* (2020), 495468.

[5] BEYAN, E. V. P., ROSSY, A. G. C., ET AL. A review of ai image generator: influences, challenges, and future prospects for architectural field. *Journal of Artificial Intelligence in Architecture 2*, 1 (2023), 53–65.

[6] CAO, Y., ABDUL AZIZ, A., AND MOHD ARSHARD, W. N. R. Stable diffusion in architectural design: Closing doors or opening new horizons? *International Journal of Architectural Computing* (2024), 14780771241270257.

[7] CASTELO-BRANCO, R., AND LEITÃO, A. Algorithmic design in virtual reality. *Architecture 2*, 1 (2022), 31–52.

[8] CHEUNG, L. H., AND DALL'ASTA, J. C. Exploring a collaborative and intuitive framework for combined application of ai art generation tools in architectural design process. In *AISB Convention 2023 Swansea University 13/14 April 2023* (2023), p. 122.

[9] CHEUNG, L. H., AND DALL'ASTA, J. C. Human-computer interaction (hci): Approach to artificial intelligence in education (aied) in architectural design. *Eídos 17*, 23 (2024), 109–131.

[10] CHO, J. Y., AND SUH, J. Spatial ability performance in interior design and architecture: Comparison of static and virtual reality modes. *Buildings 13*, 12 (2023), 3128.

[11] DOMA, O. O., AND ŞENER, S. M. Dreamscape bricks vr: An experimental virtual reality tool for architectural design. *Interaction Design and Architecture (s) 52*, 52 (2022), 234–258.

[12] DOMA, O. O., AND ŞENER, S. M. An investigation of architectural design process in physical medium and vr. *A| Z ITU JOURNAL OF THE FACULTY OF ARCHITECTURE 19*, 3 (2022), 631–649.

[13] DROGEMULLER, A., SAKHAEI, H., CUNNINGHAM, A., YU, R., GU, N., AND THOMAS, B. H. Envisioning paramersive design: An immersive approach to architectural design and review. In *2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)* (2023), IEEE, pp. 261–265.

[14] ELIANA, S., FAYOLA, K., MARCELINO, B., PAMUDJI, M., AND NURADHI, M. Development of ai and vr technology in architectural firms. *Aksen: Journal of Design and Creative Industry 8*, 3 (2024).

[15] EMORY, M., AND DOSS, W. How tall is that mushroom? design and story telling with artificial intelligence, virtual, and augmented reality. In *EAEA16: Envisioning Architectural Scales in the Analogue and Virtual Representation of Architecture* (2023), Royal Danish Academy-Architecture, Design, Conservation, pp. 325–332.

[16] ENGELI, M., AND KURMANN, D. A virtual reality design environment with intelligent objects and autonomous agents. *Proc of Design and Decision Support Systems* (1996).

[17] ERGÜN, O., AKIN, Ş., DINO, İ. G., AND SURER, E. Architectural design in virtual reality and mixed reality environments: A comparative analysis. In *2019 IEEE conference on virtual reality and 3D user interfaces (VR)* (2019), IEEE, pp. 914–915.

[18] HSU, T.-W., TSAI, M.-H., BABU, S. V., HSU, P.-H., CHANG, H.-M., LIN, W.-C., AND CHUANG, J.-H. Design and initial evaluation of a vr based immersive and interactive architectural design discussion system. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (2020), IEEE, pp. 363–371.

[19] HUSSAIN, R., SABIR, A., LEE, D.-Y., ZAIDI, S. F. A., PEDRO, A., ABBAS, M. S., AND PARK, C. Conversational ai-based vr system to improve construction safety training of migrant workers. *Automation in Construction 160* (2024), 105315.

[20] IBRAHIM, M., AND EL SHAKHS, A. Virtual reality as a design tool to achieve abstract concepts of spatial experience: A case study of design studio teaching. *Civil Engineering and Architecture 11*, 2 (2023), 1010–1019.

[21] KADRY, A., AND HUSSIEN, E. Applications of virtual reality technologies in the field of design and arts. *International Journal of Multidisciplinary Studies in Art and Technology 6*, 2 (2023), 44–70.

[22] LEE, J.-K., LEE, S., KIM, Y.-C., KIM, S., AND HONG, S.-W. Augmented virtual reality and 360 spatial visualization for supporting user-engaged design. *Journal of computational design and engineering 10*, 3 (2023), 1047–1059.

[23] LI, C., ZHANG, T., DU, X., ZHANG, Y., AND XIE, H. Generative ai models for different steps in architectural design: A literature review, 2024.

[24] M MATTER, N., AND G GADO, N. Artificial intelligence in architecture: Integration into architectural design process. *Engineering Research Journal 181* (2024), 1–16.

[25] MENG, Q., GE, M., AND FENG, Z. The integration of artificial intelligence in architectural visualization enhances augmented realism and interactivity. *Academic Journal of Science and Technology 12*, 2 (2024), 7–12.

[26] OF ARCHITECTS (AIA), T. A. I. The architect's journey to specification: Artificial intelligence adoption in architecture firms – opportunities & risks. *AIA* (2024).

[27] ONATAYO, D., ONOSOSEN, A., OYEDIRAN, A., OYEDIRAN, H., AROWOIYA, V., AND ONATAYO, E. Generative ai applications in architecture, engineering, and construction: Trends, implications for practice, education & imperatives for upskilling—a review. *Architecture 4* (10 2024), 877–902.

[28] PLOENNIGS, J., AND BERGER, M. Ai art in architecture. *AI in Civil Engineering 2*, 1 (2023), 8.

[29] ROUPÉ, M., JOHANSSON, M., MAFTEI, L., LUNDSTEDT, R., AND VIKLUND-TALLGREN, M. Virtual collaborative design environment: Supporting seamless integration of multitouch table and immersive vr. *Journal of Construction Engineering and Management 146*, 12 (2020), 04020132.

[30] SHEHADEH, A., AND ALSHBOUL, O. Enhancing engineering and architectural design through virtual reality and machine learning integration. *Buildings 15*, 3 (2025), 328.

[31] SHI, Y., GAO, T., JIAO, X., AND CAO, N. Understanding design collaboration between designers and artificial intelligence: A systematic literature review. *Proceedings of the ACM on Human-Computer Interaction 7*, CSCW2 (2023), 1–35.

[32] URBAN DAVIS, J., ANDERSON, F., STROETZEL, M., GROSSMAN, T., AND FITZMAURICE, G. Designing co-creative ai for virtual environments. In *Proceedings of the 13th Conference on Creativity and Cognition* (2021), pp. 1–11.

[33] VERVERIDIS, D., NIKOLOPOULOS, S., AND KOMPATSIARIS, I. A review of collaborative virtual reality systems for the architecture, engineering, and construction industry. *Architecture 2*, 3 (2022), 476–496.

[34] WANG, A., DONG, J., LEE, L.-H., SHEN, J., AND HUI, P. Towards ai-architecture liberty: A comprehensive survey on design and generation of virtual architecture by deep learning, 2024.

[35] WU, S., AND HAN, S. System evaluation of artificial intelligence and virtual reality technology in the interactive design of interior decoration. *Applied Sciences 13*, 10 (2023), 6272.

[36] YU, R., GU, N., LEE, G., AND KHAN, A. A systematic review of architectural design collaboration in immersive virtual environments. *Designs 6*, 5 (2022), 93.

[37] YU, R., GU, N., LEE, G., AND KHAN, A. A systematic review of architectural design collaboration in immersive virtual environments. *Designs 6*, 5 (2022), 93.

[38] ZEYTIN, E., KÖSENCIĞ, K. Ö., AND ÖNER, D. The role of ai design assistance on the architectural design process: An empirical research with novice designers. *Journal of Computational Design 5*, 1 (2024), 1–30.

[39] ZHANG, C., WANG, W., PANGARO, P., MARTELARO, N., AND BYRNE, D. Generative image ai using design sketches as input: Opportunities and challenges. In *Proceedings of the 15th Conference on Creativity and Cognition* (2023), pp. 254–261.