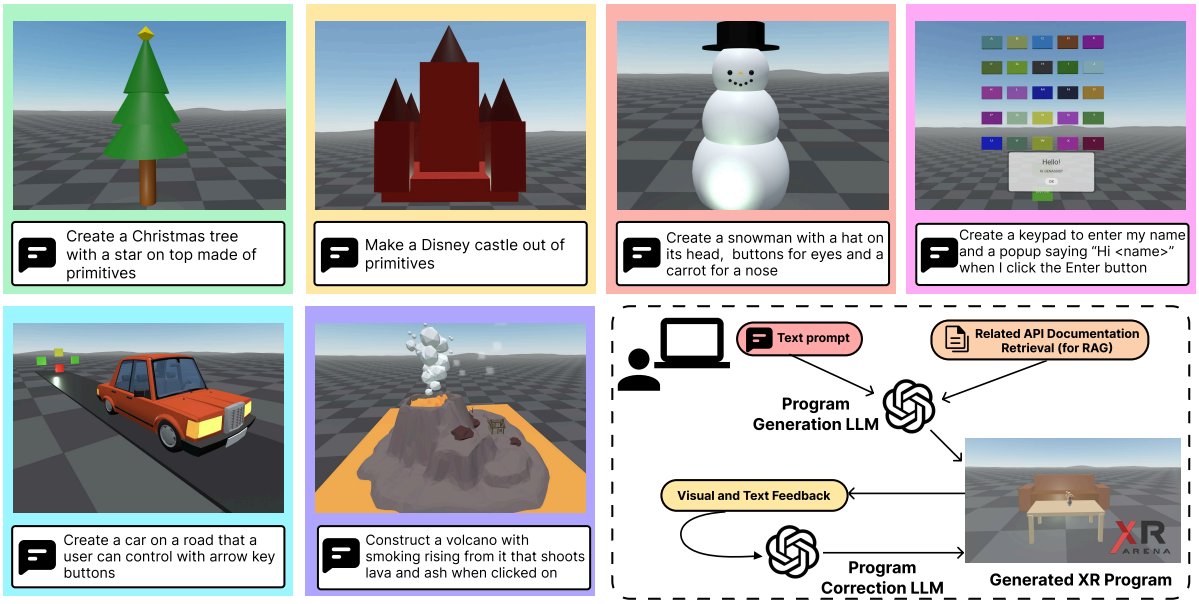
GenAssist: Interactive Prompt-Driven XR Program Generation



(a) Example Programs Generated using GenAssist **(b) High Level System Diagram and Workflow for GenAssist**

Figure 1: **GenAssist overview and examples. (a)** Example XR programs generated from natural-language prompts: procedural modeling with a set of base objects ("Christmas tree with a star"; "fantasy castle"), compositional object creation ("snowman with a hat, button eyes, and a carrot nose"), GUI and event handling ("keypad to enter a name and a pop-up saying 'Hi <name>' on Enter"), continuous control ("car that the user drives with arrow keys"), and event-triggered effects ("volcano that smokes and erupts on click"). *Images are static renderings; many examples involve interactions (e.g., clicking, and animations) that are present in the generated programs but not visible in the stills.* **(b)** System workflow: a user writes a text prompt; a Program-Generation LLM (aided by retrieval-augmented API documentation) synthesizes code and generate an XR program; the system then captures visual and textual snapshots; a Program-Correction LLM inspects these artifacts and edits the code; the loop repeats until the requested scene and behavior are achieved, yielding the final XR program.

Abstract

This paper introduces GenAssist, a system for generating interactive Extended Reality (XR) programs from natural language prompts. Given plain text descriptions of desired programs, our system uses Retrieval-Augmented Generation (RAG) to retrieve related documentation and example code, which are then used to prompt Large Language Models (LLMs) to generate and execute hot-pluggable XR programs in real time. To ensure that the programs are written correctly to the user's specifications, we add a closed-loop feedback mechanism using virtual cameras in the scene that iteratively refines the system's output, mimicking the development cycle of human developers that compile and then interactively test programs. GenAssist generates scripts that can not only place multiple primitives and 3D models in plausible locations in a virtual scene, but it can also animate and enable user interactions with those objects. We show that across a benchmark of 50 diverse XR program prompts, our system achieves high output accuracy and program generation quality. Furthermore, we conduct a user study with 10 participants that demonstrates GenAssist's effectiveness and usability (NASA TLX = 39) for XR program generation. We compare GenAssist to prior systems and show that it is significantly faster (<10 seconds per run) and makes fewer LLM calls.

Index Terms: Virtual Reality, Extended Reality, Large Language Models, Program Generation

1 Introduction

With the rapid advancement of Extended Reality (XR) technologies, 3D applications will one day move beyond niche use cases and become part of everyday experiences. Once limited to specialized fields, 3D content will soon be found in a wider range of applications, from immersive entertainment to interactive tools that blend digital content with the physical world. Despite this growing adoption, the development of XR programs remains complex. It often requires a steep learning curve and a combination of skills in design, programming, art, and many software tools. Even creating simple 3D interactions can be time-consuming, while advanced applications require significant effort and expertise. This high baseline effort creates a barrier for non-experts, limiting opportunities for rapid prototyping, experimentation, and broader participation in XR content creation.

Similarly, AI-powered chatbots and assistants are rapidly advancing to the point where they can be used as a unique tool to simplify XR authoring. While this vision is promising, text-only chatbots are not yet sufficient for highly visual workflows like XR development, which rely on iterative, real-time visual feedback. Ideally, we need mechanisms that can both generate XR programs and give developers visual feedback into the scene so that they can iteratively evaluate their output.

This paper presents **GenAssist**, a system that enables users to create interactive XR programs in real time using natural language prompts. GenAssist leverages Large Language Models (LLMs) to generate code that integrates with interactive XR scripting environments. In our implementation, we target programs that run on the ARENA platform [31], which exposes a WebXR front-end to dynamically load and execute Python programs. Users can interact with these programs in standard browsers (in 3D) or in immersive mode on headsets. GenAssist creates programs that allow users to place 3D objects, import existing models, create animations, and define interactive behaviors, all by simply describing their goals in natural language. For example, a user can simply enter *"create a tree"* instead of manually searching for a model or assembling one from primitive objects, or *"make the cube rotate when clicked"* rather than writing event-driven code from scratch. Our system also supports iterative program development, allowing users to modify and expand XR programs as their goals evolve.

GenAssist is designed to democratize XR content creation by enabling non-experts to generate small, interactive 3D programs purely using natural language. Rather than replacing professional XR development tools, our system targets rapid prototyping, educational content creation, and experimentation scenarios where ease of use and quick iteration are more important than scene complexity or performance.

To ensure the accuracy and reliability of generated XR code using GenAssist, we incorporate two key techniques:

First, GenAssist employs a **self-correction feedback mechanism** that helps to ensure that the generated XR program matches the intent of the user. The system continuously evaluates the generated program using visual feedback, spatial information, and the state of the program to identify discrepancies between what the user requested and what was produced. When misalignments or inaccuracies are detected, GenAssist automatically refines previously generated code to bring the output closer to the intended result. This feedback loop supports more accurate and robust content generation over time.

Second, it pulls semantically relevant information from platform-specific documentation and relevant code examples using **Retrieval-Augmented Generation (RAG)** to ground LLM outputs. This not only ensures that generated programs are relevant to the user's query but also prevents syntactical and potential runtime errors by providing the LLM with context tailored to the target runtime environment and task. This improves the accuracy of the generated code especially in generating ARENA-specific code, which is a rapidly evolving and improving platform.

To evaluate GenAssist, we assess its accuracy and performance in generating a diverse set of XR programs. Specifically, we propose a benchmark of 50 programs of varying complexity, covering object placement, animations, and user interactions, which we use to test the XR output of GenAssist. We compare GenAssist against several baselines, including ablated variants without the feedback loop or retrieval module, as well as previous work on prompt-based XR program generation. Since evaluating an XR program is highly subjective and there could be multiple correct programs, we have human evaluators rate the accuracy and quality of outputs. GenAssist achieves the highest average rating in all metrics, outperforming standard GPT-4o and other baselines, highlighting the value of our feedback and retrieval mechanisms.

We validate our system through a structured user study with 10 participants, measuring the user experience during XR program creation with our system. The system received a NASA TLX score of 39.0/100 indicating low perceived workload and a system usability score (SUS) of 69 indicating good usability.

Finally, GenAssist shows strong efficiency, with an average generation time of 9.93 seconds per query and 14.17 seconds per correction. It also requires significantly fewer LLM calls per scene compared to prior systems while producing more accurate results.

In summary, our paper contributes the following:

1. **GenAssist**: an open-source system for generating XR programs from natural language prompts, enabling rapid 3D scene creation and interaction design.[1]
2. A visual feedback loop for LLM generated code refinement to improve the quality of XR programs generated.
3. A user study and system-level analysis that provides information on GenAssist's usability, generation accuracy, and runtime performance compared to existing approaches.
4. A comprehensive evaluation of our system's program generation quality compared to state-of-the-art baselines across 50 diverse prompts, with all prompts and outputs released as a public benchmark dataset.

2 Related Work

2.1 XR Prototyping and Generation

To create 3D scenes and interactive programs, designers have traditionally relied on commercial game engines such as Unity [38] and Unreal Engine [11], which include plugins like MRTK [27] which provides primitives for 3D user interfaces. These platforms often have a high barrier to entry, as developing advanced applications requires skilled developers or experienced game designers. This makes rapid prototyping of 3D scenes or interactive content difficult and time-consuming for non-experts who may only need a small application for a quick 3D visualization. As a result, there is growing interest in democratizing the process of 3D content creation, allowing users to build interactive environments without requiring extensive technical expertise.

Early approaches to the generation of 3D and XR content focused on scene adaptation using predefined rules, semantic reasoning, and domain-specific heuristics. These systems typically modified existing scenes based on spatial or contextual cues rather than generating content from scratch [26, 25, 47]. For example, Cheng et al.'s *SemanticAdapt* [8] adjusts object layouts based on semantic relationships, while Chang et al.'s *SceneSeer* [7] allows natural language prompting to search through existing scenes and models and place them in the scene. Other systems like Xu et al.'s *Sketch2Scene* [43] allow users to sketch 3D scenes, relying on model training and visual templates to construct environments. Adobe's Aero [2] enables interaction with AR elements through a GUI, reducing the need for programming. Although these systems provide more intuitive authoring tools than scripting in a game engine, they typically only support static scenes and have limited ability to generalize across domains. In addition, they lack support for open-ended user interactions with the scene objects.

With the rapid growth of large language models and generative AI, emerging systems have begun to support text-driven 3D scene creation, providing a more natural interaction modality compared to scripting or graphical user interfaces. Diffusion-based methods [33, 24, 20, 15] generate 3D assets from natural language, focusing primarily on mesh generation. Though powerful for object synthesis, these approaches are typically limited to rigid, static models and do not extend to interactive or programmable XR content. Similarly, Google has been releasing a series of world building models such as Genie [12] that can create high-fidelity 3D worlds

[1] The code will be open-sourced at camera ready.

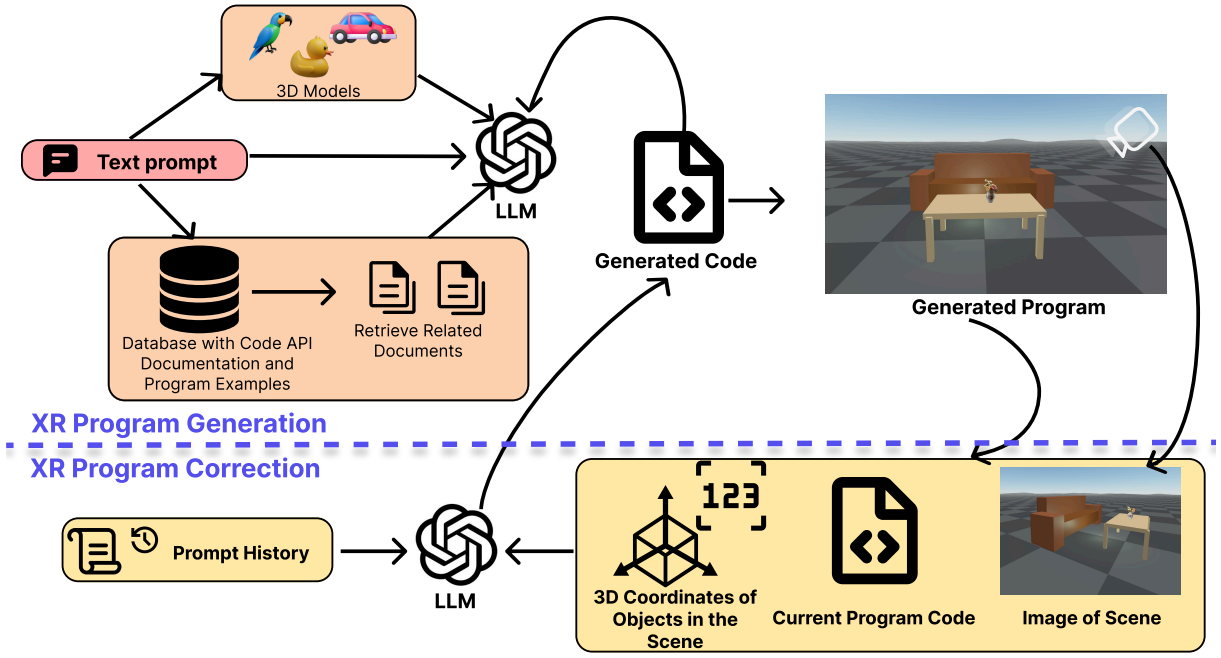


Figure 2: **GenAssist system architecture.** GenAssist comprises of two main stages: program generation and program correction. In the generation stage, a user-provided text prompt is used to retrieve relevant ARENA platform documentation, API usage examples, and 3D models using RAG techniques. These are passed to a large language model (LLM) to generate code, which is then executed in ARENA's Python runtime to produce the XR program. In the correction stage, the system gathers the current program code, 3D object coordinates, and images of the scene from strategically-placed virtual cameras, along with the user's prompt history. This context is used by the LLM to make code corrections if needed.

and have basic agent interaction, but are still mostly static and only support limited interactions.

Scene generation and editing systems using LLMs have also been rapidly advancing. For instance, Qian et al. introduced *SHAPE-IT* [34], which leverages LLMs to translate user prompts into code that generates shapes on pin-based shape displays. Closely related to GenAssist are systems such as *3D-GPT* [37] and *BlenderAlchemy* [17], both of which generate Blender-compatible Python code to synthesize and edit existing geometry and materials. *3D-GPT* emphasizes reasoning and task decomposition from textual instructions, whereas *BlenderAlchemy* integrates visual processing capabilities, enabling the LLM to interpret both text and image inputs for a more interactive and multimodal approach to 3D content generation. Building on these efforts, *SceneCraft* [16] introduces a self-correction mechanism that iteratively refines generated scenes to address errors. GenAssist builds upon the ideas presented in these works to create *interactable* scenes using existing 3D models and object primitives. Specifically, we make use of a scripting API and runtime that provides straightforward and accessible methods for scene manipulation that an LLM can leverage.

Although LLMs have been applied to XR content generation, many existing approaches are tailored to highly specific use cases rather than general-purpose scene creation. Applications such as *VR Copilot* [48] and *Holodeck* [44] leverage LLMs to specifically generate room layouts in Unity. Other applications explore this in the context of video games by dynamically generating in-game objects using text or player inputs [19, 35]. These systems show the potential of LLMs for 3D scene modification, but lack the adaptability needed for general-purpose, open-ended XR creation.

In the domain of interactive XR program generation, multiple systems have leveraged LLMs to create interactable content. Giunchi et al. with *DreamCodeVR* [13] present a simpler approach to generating objects, performing a LLM call to generate code for object creation. However, it lacks an iterative refinement process and exhibits limited complexity in object outputs. De La Torre et al.'s *LLMR* [10], a recent system for XR program generation, employs a pipeline of multiple LLM calls to plan, analyze, build, and refine scene generation outputs, which we discuss in Section 4.2 as a recent example of multistage XR program synthesis using LLMs. Conceptually, our approach aligns with broader efforts like PAIL [46], which reframe LLM-based programming as a design activity involving iterative exploration and decision tracking.

2.2 RAG for Code Generation

RAG enhances the capabilities of LLMs by supplementing their internal knowledge with relevant external context. Initially developed to improve accuracy and reduce hallucinations in natural language tasks [22], RAG retrieves relevant documents or code snippets from a corpus and incorporates them into the generation process. This is especially useful in domains where training data alone may be insufficient, enabling dynamic access to up-to-date or domain-specific information. While RAG is widely used in natural language processing, it has also been adapted for multimodal tasks, such as image generation with retrieval-augmented diffusion models [3] and 3D content generation [36].

In code generation, RAG has shown promise but also presents challenges. Wang et al. introduce *CodeRAG-Bench* [40], a benchmark evaluating RAG-based code generation, noting that while retrieved context can improve results, models often struggle to integrate semantically relevant but lexically mismatched content. Large-scale systems such as *CodeRetriever* [23] demonstrate the effectiveness of unimodal and multimodal retrieval strategies for code synthesis and search. Systems such as *REDCODER* [30] retrieve and incorporate various code–text pairs, and others have ex-

Program Generation Prompt=
You are a 3D XR program generation assistant. Based on the following context, generate code to create the specified program:

1. Relevant Documentation and Examples:
{retrieved_docs_and_examples}

2. Available 3D Models:
{model_urls}

3. Current Scripts that are Running:
{current_running_scripts}

4. History of prompts (first one is the most recent):
{prompt_history}

Generate the code considering the above context for the question that follows: **{question}**

(a) Program Generation Prompt

Program Correction Prompt=
Based on the following context, decide whether to modify the program or not. If you decide to modify it, provide the python script of the corrected program:

{image_of_scene.png}

Current Script Running in the scene:
{current_running_scripts}

Scene Object Bounding Boxes [format = objectid: min: minimum xyz coordinate, max: maximum xyz coordinate] (if the bounding box is inf that means the model is bad and should be replaced):
{bounding_boxes}

History of Prompts(first prompt is the most recent):
{prompt_history}

Use the image and the information to correct the program.

(b) Program Correction Prompt

Figure 3: Template prompts used to provide additional context for program generation and program correction.

plored the retrieval for specialized domains, including the generation of RTL codes for hardware design [21, 41], the enhancement of code security [29] and automated bug fixing [39]. However, studies show that irrelevant retrieval can negatively impact performance, highlighting the importance of high-quality retrieval strategies [45].

Building on these foundations, GenAssist applies RAG to XR program synthesis, a novel extension of retrieval-augmented code generation to 3D and interactive environments.

3 System Design

In the following subsections, we outline how GenAssist generates XR programs from natural language for ARENA (a WebXR runtime platform), how it employs a feedback loop to iteratively refine its output, and how retrieval-augmented generation (RAG) techniques are applied to enhance code generation. An overview of the system architecture is shown in Fig. 2, with descriptions of each component described below.

3.1 Code Generation for ARENA

ARENA provides a dedicated Python API, `arena-py` [42], and a Python runtime. All program I/O passes through a scene graph that is overlaid on a pub-sub backend. The choice of Python is particularly advantageous: It is one of the most widely used programming languages with a vast ecosystem of publicly available code. Thus, most commercial large language models (LLMs) such as GPT-4o [1] have been trained on large amounts of Python programming data, making these models well suited to generate syntactically correct and directly executable Python code.

Although ARENA is primarily web-based, its programs are inherently cross-platform and distributed, meaning they can run on any device with a network connection and are not limited to a specific viewing environment. In fact, all our experiments were conducted using a browser-based viewing app, though ARENA also supports a Unity-based viewer. It is important to note that while our implementation targets ARENA, the core methodology generalizes to any XR platform. The visual feedback loop requires only the ability to capture screenshots and extract object spatial information, capabilities available in Unity, Unreal Engine, and other platforms through their respective APIs. Similarly, our RAG approach can incorporate documentation and examples from any platform's ecosystem. ARENA was selected primarily for its hot-pluggable execution model, which facilitates rapid iteration during development and evaluation, as well as structured and easy-to-access documentation.

GenAssist *only* synthesizes the `arena-py` code, the ARENA runtime handles event dispatch, scene synchronization, and execution. As of this submission, the interactions exposed through ARENA to `arena-py`, and hence the interactions that GenAssist can use in the programs it generates, consists of (a) pointer/cursor events (mouse or controller interactions), (b) proximity-based interaction, and (c) a limited set of keyboard events. This is a platform boundary rather than a limitation of our method, and so on runtimes with richer inputs, GenAssist can target those events as long as documentation and examples are available for retrieval.

3.2 Iterative Scene Correction Feedback Loop

Although LLMs have demonstrated strong capabilities in generating code across a wide range of domains, they are not error-free. LLMs can still produce syntactically invalid code, incorrectly use APIs, or introduce logical or physical inconsistencies, which affects the quality of generated XR programs. To mitigate these issues and improve reliability, GenAssist incorporates a visual and textual feedback loop inspired by how developers typically refine XR applications: writing code, observing the resulting scene, and iteratively adjusting based on visual output as seen in Fig. 2 XR program correction.

To replicate this workflow, GenAssist periodically captures 2D screenshots of the scene, taken from strategic camera positions that provide a comprehensive view of all objects in the environment. To capture these images, we render the ARENA scene in a browser window using Playwright [28], a browser automation framework, and take screenshots of the generated program. To guarantee that the entire scene is captured in the image, the system automatically computes the 3D bounding boxes of all objects and determines an appropriate camera position that ensures that all objects are within view. This visual feedback allows the model to "see" what has been generated, enabling it to detect issues that may not be obvious from the code alone. This process can be executed on a separate machine, ensuring that the user's performance remains unaffected.

Additionally, GenAssist feeds the 3D bounding box coordinates of all the objects to the program correction module. This helps the LLM place objects in physically plausible locations by providing an understanding of spatial relationships within the program, which are not always evident from 2D screenshots, such as relative object sizes and placements. This is especially important when incorporating external objects and models in ARENA, as the generated code only references a file server URL (e.g., `https://arenaxr.org/fakeuser/mymodels/model.glb`) without specifying the underlying geometry. As a result, the system cannot determine the model's relative size or shape until it has been rendered in the scene. This spatial information helps the LLM understand a given 3D model's positioning, scale, and potential overlaps with other objects, providing essential context for reasoning about spatial relationships within the 3D environment. Alongside visual and spatial feedback, the loop also includes the current generated code and a full history of previous prompts to

ensure that the model remains aware of both the intended goals and the program's evolution over time.

Lastly, to improve robustness, the system captures and feeds any detected errors, including syntax issues, runtime exceptions, or execution logs, into the correction loop. This loop iteratively refines the scene by prompting the LLM to fix identified problems, correct object placements, and adjust the program to better align with the intended design. Through this continuous cycle of feedback and correction, GenAssist ensures that the generated XR programs become increasingly accurate and functional over time. The template for the data provided as the context for the scene corrector is also shown in Fig. 3. This is one of the key features that allows for iterative development, where a person's prompt can reference previous actions, the current scene, and program state, building up and modifying a program using multiple sequential prompts.

3.3 XR Program Generation with Retrieval-Augmented Generation

To enable GenAssist to generate code that follows ARENA's syntax, structure, and feature set, we supplement the LLM with additional context in the form of curated examples and API documentation. Although LLMs are powerful, it is well known that they can hallucinate and produce incorrect responses, especially when they lack up-to-date or domain-specific knowledge [18]. This is especially true for rapidly evolving platforms like ARENA, where APIs and conventions frequently change, or for complex ecosystems like Unity and Unreal Engine, where the breadth of features and multiple engine versions can make it difficult for the LLM to determine which details are most relevant to a given question. Previous work like LLMR [10] handled this using another LLM call to identify what information needs to be provided as a context, which can be expensive and time consuming. To address this, we used a lightweight RAG-based approach to improve the model's ability to produce accurate and relevant code. At a high level, GenAssist generates code by prompting the LLM with a selected blend of documentation, usage examples, and the current state of the program.

A key component of this process is retrieving relevant content from ARENA's API documentation and example code. By scraping documentation webpages and code repositories, we create a vector database of semantic embeddings mapped to text. When a user queries GenAssist, it will use RAG to tokenize the prompt and retrieve the closest semantic matches from the database. The text of those matches is then injected into the prompt before being sent to the LLM. By including documentation and examples directly in the prompt, the system leverages in-context learning [5], which is a known strength of LLMs, allowing the model to better understand API usage patterns, coding style, and program structure. We found that this improves the accuracy of the generated code and reduces hallucinations.

Furthermore, ARENA includes a public file server of 3D models (in `.gltf/.glb` and `.obj` formats). Using a similar technique as described above, GenAssist creates a vector database using embeddings created from file metadata that can be queried using RAG. For each user query, we search the vector database to find any 3D models that might be useful for the LLM to use when generating the program, allowing for easy model integration into the system.

To maintain program continuity and context, the system also feeds the current running XR program directly into the prompt. This allows the LLM to understand what is already present in the scene and allows it to directly edit the current program. Furthermore, GenAssist appends a history of prior prompts and responses to preserve the flow of user intent and interactions over time. This running context ensures that the model's generation aligns with both the current scene and the user's iterative design process. We provide the template for the context provided to the LLM in Fig. 3.

For our system, we use GPT-4o [1] as the LLM and Chroma-db [9] as the RAG vector database. We use OpenAI's embedding models (specifically *text-embedding-ada-002*) to embed ARENA documentation and `arena-py` examples, with each page or example being an entry into the database. Similarly, we also use the same embedding model for the 3D models, which we save in a separate chroma-db database.

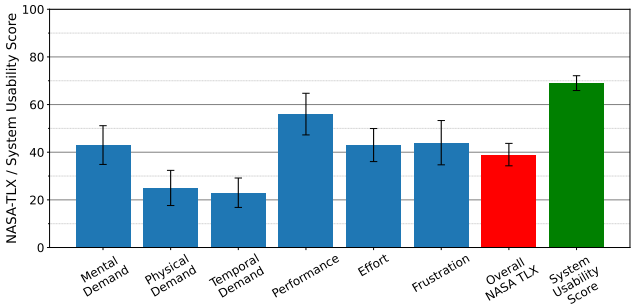


Figure 4: Average Usability Scores across all participants. A lower NASA TLX score indicates lower perceived workload (lower is better for blue and red), while a higher System Usability Score indicates better usability (higher is better for green).

4 Evaluation

To be effective, GenAssist must provide a pleasant user experience, generate plausible and intent-aligned XR programs, and operate with low overhead. We evaluate the system across these three key dimensions: (1) User Experience – Does GenAssist lower technical barriers and enable successful task completion in XR program creation? (2) Generation Quality – Does it produce XR programs that are both accurate and consistent with user intent? (3) System Overhead – What is the runtime cost of the system, measured in terms of latency and the number of LLM queries required?

4.1 User Experience

To evaluate the user experience of GenAssist, we conducted a study with 10 participants (7 male and 3 female between the ages of 18 and 55). Each participant received a brief introduction to the system and its functionalities and was given five minutes to freely explore GenAssist by generating their own programs and exploring the system's capabilities. After the familiarization phase, participants were presented with a pre-generated *target* program and given time to interact with it and explore its functionality. They were then asked to use GenAssist to reproduce the program. They were allowed to make as many prompts as they wished until they were satisfied with the result. Participants performed this task twice, each time with a different target program representing a distinct level of difficulty. Both programs were designed to evaluate how effectively users could achieve specific goals with the system. To ensure consistency, all participants attempted the same two programs, which are included in the supplementary material.

After completion of the tasks, the participants were given a short survey comprising of the NASA Task Load Index (NASA TLX) [14] and the System Usability Scale (SUS) [4]. The specific XR programs used for testing and the survey questionnaires are included in the supplementary material for reference.

Fig. 4 presents the usability results, including the six NASA TLX metrics, the overall average NASA TLX workload score, and the SUS score. Lower NASA TLX scores indicate reduced workload and effort. Our results show a NASA TLX overall score of **39.0 out of 100**, suggesting that using GenAssist imposes a relatively low cognitive load. Among NASA TLX metrics, *performance* had the highest score (indicating a higher perceived difficulty in being suc-

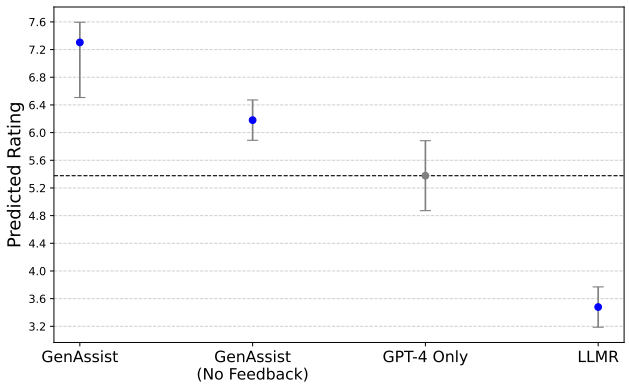


Figure 5: Estimated system effects from the linear mixed effects model with 95% confidence intervals. Coefficients represent the actual rating values predicted by the model. Higher the value, higher the rating which is out of 10.

cessful in the task), which aligns with participant feedback. Users reported that generating correct output sometimes required multiple iterations of prompting, particularly for complex programs. This iterative re-prompting process increased the perceived effort to successfully complete the task, although the participants acknowledged that this would be much easier than writing code to generate these programs.

In contrast, the System Usability Scale (SUS), where a higher score indicates more usability, yielded a score of 69 for GenAssist. With 68 generally considered average, this suggests that the system is reasonably usable. Although SUS provides a useful overall measure of perceived usability, we place greater emphasis on the NASA TLX results for this evaluation. NASA TLX captures a broader picture of user experience by measuring perceived workload across dimensions such as mental demand, effort, and frustration. This makes it particularly well-suited for evaluating systems like GenAssist, where user effort can be affected not just by the interface but also by the behavior of the underlying model. For example, instances where users needed to re-prompt the system to refine outputs contributed directly to higher perceived workload. As such, the NASA TLX offers a more comprehensive understanding of how challenging or demanding the system is to use in practice.

4.2 Generation Accuracy

Our goal is to evaluate the accuracy of GenAssist in generating XR programs. However, defining what constitutes an "accuracate" or a "good" output in this context is inherently subjective. For instance, a prompt such as "make a car" could be satisfied by a simple 3D car model, by a basic construction using a rectangular cube with four cylinders as wheels, or by a more detailed, higher-fidelity car composed of primitives with windows, realistic proportions, and additional features.

While prior work has reported accuracy and error metrics, the underlying evaluation methods can be unclear or insufficiently detailed. To address this gap, we conducted a study in which participants, referred to here as raters, evaluated the correctness of XR programs generated from their prompts. Each program was assessed along four distinct metrics, described below, using a 10-point scale.

Program Correctness Metrics:

1. **Prompt Match**: How closely does the XR program align with the input prompt?
2. **Object Placement**: How well are the objects positioned within the scene?
3. **Functionality**: Does the program behave and function as expected?
4. **Overall Quality**: What is the overall perceived quality of the experience?

For comparison, we compare four text-to-scene systems, including GenAssist: (1) **GenAssist**, (2) **GenAssist (No Feedback, RAG Only)**, (3) **GPT-4o only (No Feedback, No RAG)**, and (4) **LLMR** [10], a recent state-of-the-art system for Unity program generation.

We curate a set of 50 prompts across five categories of increasing complexities. We create a taxonomy of prompts based on the skills and techniques we expect a non-expert user to use when generating simple XR programs. Each category consists of 10 prompts. Some are drawn directly from the LLMR evaluation set, while others were newly created to reflect practical tasks users might attempt or to highlight the capabilities of the ARENA platform. The full set of prompts is provided in the supplementary material. This taxonomy consists of: (1) **Object Placement** (2) **Animations** (3) **Interactivity**, (4) **Complex Programs Combining Multiple Features**, and (5) **Iterative Program Generation**.

For the study, we recruited 15 raters with varying levels of familiarity with XR. Each rater evaluated the output of 40 prompts (distributed across the 4 systems and 5 program categories), with each prompt's output assessed by three different raters. Prompt assignments were randomized: some raters evaluated the same prompt across all systems, while others reviewed different prompts and systems. This design was intended to mitigate potential bias and ensure a diversity of scoring perspectives.

Given multiple sources of variability (system, prompt, metric, rater), we use a linear mixed-effects model [32] to estimate system effects while controlling for prompt difficulty and rater stringency. Systems and the program correctness metric are modeled as fixed effects while prompts and raters receive random intercepts, controlling for confounding factors like prompt difficulty or rater bias. This lets us attribute differences in ratings to the systems and metrics rather than variations in specific prompts or raters. We define our model as follows:

$$Rating \sim System + Metric \tag{1}$$

This is broken down into more detail:

$$Rating_{i,j,k} = \beta_0 + \beta_{\text{System}} \cdot System_{i,j,k} + \beta_{\text{Metric}} \cdot Metric_{i,j,k} + u_i + v_k + \varepsilon_{i,j,k} \tag{2}$$

Where:

i, j, k : Individual rater index, Rater ID, prompt index.
$\beta_0 = 5.377$: Intercept term (baseline for *GPT-4o Only* system).
β_{System} : Fixed effect for system type:

- 7.304 for GenAssist ($p < 0.001$).
- 6.180 for GenAssist (No Feedback) ($p < 0.001$).
- 3.479 for LLMR ($p < 0.001$).

β_{Metric} : Fixed effect for metric values.
u_i, v_k : Rater-level and Prompt-level random intercept.
$\varepsilon_{i,j,k}$: Residual error.

The baseline condition in our mixed-effects model corresponds to the GPT-4o Only system, representing GenAssist without retrieval or feedback. As shown in Fig. 5, the full GenAssist system outperforms its ablated variants and baselines, achieving a predicted rating of 7.304, compared to 6.180 for GenAssist without feedback, 5.377 for GPT-4o, and 3.479 for LLMR, even after accounting for variability across raters and prompts. These results provide strong evidence that the closed-loop GenAssist system produces more reliable and higher-quality outputs. In particular, they underscore the effectiveness of incorporating a feedback mechanism to identify and correct hallucinations, as well as the use of

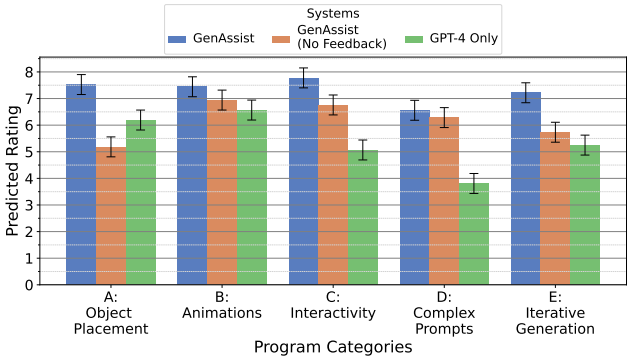


Figure 6: Accuracy of outputs across prompt categories for each system. Reported values are z-scores, obtained by normalizing ratings within each rater.

System Components	*Average Time Taken (s)*
Program Generation	**9.928**
Retrieval of Documentation and Examples	4.414
Retreival of 3D Models	0.478
Queries to LLM (GPT-4)	5.036
Program Correction	**14.173**
Get bounding boxes of objects in the scene	0.152
Get screenshot of the scene	3.792
Queries to LLM (GPT-4)	10.229

Table 1: Average time taken for the Program Generation and Program Correction stages.

RAG to enhance robustness across a wide range of task complexities.

We acknowledge that comparisons between GenAssist and LLMR are not direct for several reasons. First, LLMR was developed for Unity C#, whereas GenAssist generates Python code for ARENA. While Unity is more prevalent in XR development, Python is more widely represented in LLM training data overall, creating different baseline capabilities. Second, LLMR uses GPT-4 while our system leverages GPT-4o. Updating LLMR to use GPT-4o would require non-trivial modifications to its multi-stage prompt pipeline and Unity-specific examples, potentially compromising the integrity of the original system design. Therefore, we include LLMR as a contextual comparison to generally understand where our system stands compared to prior work, rather than a statistically significant performance claim. Our core contributions—specifically, the effectiveness of RAG and visual feedback—are validated through ablation studies on GenAssist variants, which enable controlled comparisons under identical infrastructure.

To further investigate how the performance of the different systems vary across prompt categories, we model the data with another mixed-effects model specified as follows:

$$Rating \sim System \times Category + Metric \tag{3}$$

This model allows us to gain insight on how the ratings vary for different systems for the different prompt categories A through E, which are shown in Fig. 6. We exclude LLMR from this category-wise analysis as it serves as a contextual comparison rather than a direct baseline, with the overall performance comparison shown above in Fig. 5 being sufficient to situate our work. Across all categories, GenAssist consistently outperforms its ablated variants. Interestingly, for simpler prompts (Category A), the RAG-only variant, *GenAssist (No Feedback, RAG Only)*, underperforms. Upon closer analysis, we found that this was often due to irrelevant retrievals. For simpler tasks, the language model and the few-shot examples in the prompt are usually sufficient to generate correct code. However, when extraneous documents are retrieved, the model tends to over-prioritize them, leading to hallucinated or incorrect outputs. This issue stems from the model's strong bias toward its immediate context [6, 45], causing it to pay attention to less relevant information even when simpler logic would suffice.

Thus, in isolation, RAG may not be as helpful and can even hinder performance in basic, straightforward tasks like category A. However, when combined with closed-loop feedback, the system can identify and recover from these errors, preserving strong performance even on simple prompts. However, for more complex prompts, the additional retrieved context becomes significantly more useful, as it provides the model with examples or patterns that may not be covered by the initial prompt itself. This demonstrates that retrieval is particularly effective when prompt complexity increases, but must be paired with mechanisms like feedback to ensure robustness across the full range of task difficulty.

4.3 System Overhead

To evaluate the run-time cost of GenAssist, we measured both system latency and the average number of LLM queries required during program generation and correction. These factors directly impact the system's responsiveness and determine its practicality for interactive use.

Each generation cycle introduces several sources of latency. For every user prompt, the system performs a retrieval of relevant documentation and example code, as well as 3D models to construct the LLM prompt context. Although this step adds some delay, the most significant overhead comes from the LLM response time during code generation. Once generated, running the program has negligible overhead due to ARENA's hot-pluggable execution model. The scene correction loop also introduces additional latency. Capturing scene screenshots and extracting 3D bounding boxes of all the objects in the scene adds runtime costs. However, as with the generation cycle, the LLM response time is the dominant factor in this phase.

Since the number of generation and correction cycles needed vary based on prompt complexity and the number of times the corrector needs to be called, we report these costs separately. Specifically, we measure (1) the average time taken for initial program generation, (2) the average time per correction cycle, and (3) the average number of LLM queries required per program.

Our analyses focuses on four of the five prompt categories used in the generation quality evaluation in Sec. 4.2. We exclude the *Iterative Scene Generation* category because it involves multiple user inputs by design, making it difficult to compare directly with the other categories. In iterative cases, the number of input prompts and consequently, the number of LLM calls are highly variable and task dependent.

Tab. 1 summarizes the system latency for various components of GenAssist.

Because LLM calls are the main driver of latency and cost, we report the average number of LLM calls needed to generate XR programs for the different prompt categories in Tab. 2. Simple prompts generally require only a single LLM query, with no corrections needed to produce a correct response. In contrast, more complex prompts, particularly those involving animations or interactivity, often need multiple correction cycles before a satisfactory program is generated. In addition, the number of lines in the program, which corresponds to the length of the LLM output, directly impacts generation time. Therefore, our reported values represent averages across varying levels of complexity and program lengths. The reported number of LLM queries reflects the total across both

Program Category	*Average Number of LLM Queries*
Object Placement	1.90 ± 2.07
Animations	2.10 ± 1.58
Interactivity	2.10 ± 1.22
Complex Programs Combining Multiple Features	3.10 ± 2.98

Table 2: Average number of calls to LLMs needed to generate the program for the first 4 categories of programs. The first query corresponds to the prompt from the user sent ot the program generator, and any subsequent ones refer to the queries made to the program corrector. The fifth category (*Iterative Scene Generation*) is not evaluated here as the number of prompts used as input vary based on the task.

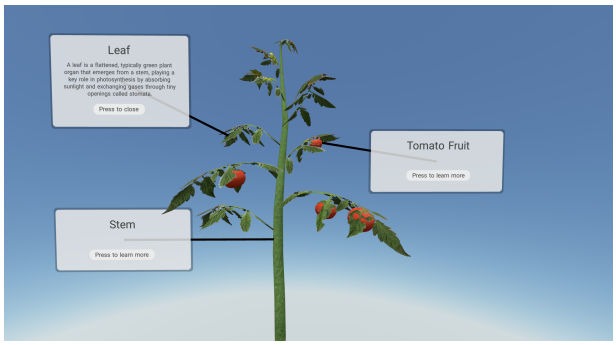


Figure 7: Example usage of GenAssist to create educational 3D programs. Here the plant is annotated with clickable cards to learn about various parts of a plant in an interactive and engaging way.

generation and correction stages. Since we exclude the *Iterative Scene Generation* category, there is only one query for program generation and any additional queries are from the scene corrector. In general, our system takes **less than 10 seconds** to generate the program code with each iteration of the corrector taking **14.2 seconds**. Overall, even under the assumption that GenAssist requires on average one LLM call for program generation and an additional call for program correction, the **total wait time of our system is 24.10 seconds, which is approximately 3.7× lower than the 90.98 seconds required by LLMR, the current state of the art.**

It is important to note that determining when a program is "complete" is inherently subjective. Our reported averages reflect the observed number of LLM queries during evaluation and serve as ballpark estimates of system overhead rather than strict upper bounds.

5 Applications

GenAssist can have a wide range of use cases, spanning both practical scenarios and open-ended creative tasks. In this section, we highlight several example applications that demonstrate how the system can be used to build interactive 3D experiences in domains such as education, remote assistance, and entertainment.

5.1 Interactive Educational Content

GenAssist can be used to create immersive and interactive educational experiences in 3D, ranging from intractable visualizations of educational concepts to more structured tools like labeled 3D models, flashcards, and interactive quizzes. For example, a biology instructor could generate a scene which includes a model of a plant with clickable parts that reveal descriptions, or a history teacher could create a virtual exhibit that students can explore. Because most educators are not experts in XR programming, tools that allow them to easily author and modify content can be extremely useful in the classroom to quickly drive up excitement. Interactivity is especially important for engagement and comprehension in learning environments, and GenAssist allows educational content to be created and adapted dynamically based on the needs of the learner. This opens the door to personalized and responsive educational tools without requiring specialized XR knowledge. An example scene can be seen in Fig. 7.

Figure 8: Example usage of GenAssist to create a animated dragon that breathes fire next to a castle.

5.2 Creative Exploration and Game Design

GenAssist enables novice users to create simple games and experiment with interactive 3D content without needing to write code. It functions as a creative playground, allowing users to explore the possibilities of XR programming through natural language prompts. One such example can be seen in Fig. 8.

5.3 Remote Assistance

In current remote assistance scenarios, such as helping someone troubleshoot equipment through VR streaming, guidance is typically limited to voice communication. However, voice alone can often be ambiguous or hard to follow, especially during complex or multi-step tasks. While some systems now support remote annotations, these features are often constrained to basic markup and lack the flexibility to add richer or more interactive content within the scene. With GenAssist, a remote assistant can quickly create 3D annotations, overlays, or even interactive elements directly within the user's XR environment. Because GenAssist operates in natural language, these can be made on the fly, making remote support sessions more effective, interactive, and intuitive. For example, during a filter-replacement procedure, the expert says: "Create a floating checklist titled 'Filter Replacement', highlight the intake valve, draw an arrow to the release latch, load `filter.glb` and place it over the target socket, add a 'Next' button that advances the checklist when clicked." GenAssist generates the ARENA code to spawn these elements and bind cursor clicks or proximity triggers to the step logic, reducing ambiguity compared to voice-only guidance.

5.4 Guided Task Assistance

GenAssist can be used to create XR programs that assist with physical-world tasks by providing contextual visual guidance. For example, it can generate annotated 3D scenes that illustrate how to operate a device or perform a step-by-step procedure. These virtual guides can include arrows, labels, and interactive elements to help users follow along more effectively. By lowering the barrier to authoring such assistants, GenAssist makes it easier to create customized instruction manuals and AR overlays without requiring technical expertise. It also enables the assistance to adapt to the user's actions for more responsive and personalized guidance.

6 Discussion

Although GenAssist enables the generation of natural language-driven XR programs with iterative refinement, it does have limitations. The system's feedback loop relies on visual observations of the current scene, specifically a rendered image, the 3D bounding boxes of all objects, and the current program code. In addition, it incorporates any run-time errors or stack traces that appear in the execution logs. This setup allows the model to detect and fix issues related to object positioning, sizing, visibility, or syntax and runtime errors. However, this feedback loop does not capture behavioral issues unless they show up as explicit errors or visible mismatches. For example, problems related to animations, event-driven logic, or interactivity, such as a button that fails to trigger an expected action, may not be detected unless they result in a scene that looks visually incorrect or raise an exception. As a result, some errors may go undetected. Addressing this would likely require integrating richer sources of runtime context, such as interaction traces or event logs.

Another challenge is reliance on embedding-based retrieval for API documentation and examples. While RAG improves LLM responses by retrieving semantically similar examples, it often retrieves documents that share keywords with the prompt, rather than those that reflect its compositional or structural intent. For example, a prompt like *"create a car and make it drive"* can lead to individual results related to cars or driving animations, but not the necessary primitives or logic to combine them into a coherent and functioning scene. Improving retrieval may require more task-specific representations or retrievers trained to support such tasks.

GenAssist targets accessibility and rapid prototyping rather than production-level XR applications. While the generated programs handle object placement, animations, and basic interactions effectively, more complex scenarios (e.g., advanced physics simulations, custom materials, or sophisticated AI behaviors) would still require traditional development approaches or further advancements in the system.

Finally, the current 3D model search approach involves a trade-off between ease of use and retrieval quality. We initially explored using the Sketchfab API for dynamic model search, but its keyword-based matching often produced results that were inconsistent or poorly suited to the prompt. To improve reliability, we instead pre-downloaded a curated set of models and indexed their metadata using embeddings. These models must then be added to the ARENA model database to be usable, which introduces setup overhead and limits flexibility. Although this process is more manual, it yielded significantly better results in practice. We will share our Sketchfab integration in our open-source code releases to support future improvements and allow the community to create alternative retrieval strategies.

7 Conclusion

We present GenAssist, a system for generating XR programs from natural language prompts by combining retrieval-augmented generation, a visual feedback loop, and hot-pluggable code execution. It allows users to iteratively build scenes that include object placement, simple animations, and basic interactivity, without writing code manually. Compared to existing systems, GenAssist achieves higher output accuracy and significantly lower latency, making it well suited for rapid XR prototyping. By reducing the technical barriers to creating and modifying XR content, it offers a practical tool for both novices and experienced developers.

References

[1] Openai gpt-4o. https://openai.com/index/hello-gpt-4o, 2024. Online. Accessed: July 2024. 4, 5

[2] Adobe. Adobe aero. Accessed: 2025-03-12. 2

[3] A. Blattmann, R. Rombach, K. Oktay, J. Müller, and B. Ommer. Retrieval-augmented diffusion models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds., *Advances in Neural Information Processing Systems*, vol. 35, pp. 15309–15324. Curran Associates, Inc., 2022. 3

[4] J. Brooke. *SUS – a quick and dirty usability scale*, pp. 189–194. 01 1996. 5

[5] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, 2020. 5

[6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, 2020. 7

[7] A. X. Chang, M. Eric, M. Savva, and C. D. Manning. Sceneseer: 3d scene design with natural language. *arXiv preprint arXiv:1703.00050*, 2017. 2

[8] Y. Cheng, Y. Yan, X. Yi, Y. Shi, and D. Lindlbauer. Semanticadapt: Optimization-based adaptation of mixed reality layouts leveraging virtual-physical semantic connections. UIST '21, p. 282–297. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3472749.3474750 2

[9] Chroma. Chroma - the open-source embedding database, 2023. Accessed: 2025-03-30. 5

[10] F. De La Torre, C. M. Fang, H. Huang, A. Banburski-Fahey, J. Amores Fernandez, and J. Lanier. Llmr: Real-time prompting of interactive worlds using large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–22, 2024. 3, 5, 6

[11] Epic Games. Unreal engine. 2

[12] P. J. B. et al. Genie 3: A new frontier for world models. 2025. 2

[13] D. Giunchi, N. Numan, E. Gatti, and A. Steed. Dreamcodevr: Towards democratizing behavior design in virtual reality with speech-driven programming. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 579–589, 2024. doi: 10.1109/VR58804.2024.00078 3

[14] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Human mental workload*, 1(3):139–183, 1988. 5

[15] L. Höllein, A. Cao, A. Owens, J. Johnson, and M. Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7909–7920, October 2023. 2

[16] Z. Hu, A. Iscen, A. Jain, T. Kipf, Y. Yue, D. A. Ross, C. Schmid, and A. Fathi. Scenecraft: An llm agent for synthesizing 3d scene as blender code, 2024. 3

[17] I. Huang, G. Yang, and L. Guibas. Blenderalchemy: Editing 3d graphics with vision-language models, 2024. 3

[18] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, Jan. 2025. doi: 10.1145/3703155 5

[19] N. Jennings, H. Wang, I. Li, J. Smith, and B. Hartmann. What's the game, then? opportunities and challenges for runtime behavior generation. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, UIST '24. Association for Computing Machinery, New York, NY, USA, 2024. doi: 10.1145/3654777.3676358 3

[20] H. Jun and A. Nichol. Shap-e: Generating conditional 3d implicit functions, 2023. 2

[21] A. Kaintura, P. R, S. S. Luar, and I. I. Almeida. Orassistant: A custom rag-based conversational assistant for openroad, 2024. 4

[22] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-

augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020. 3

[23] X. Li, Y. Gong, Y. Shen, X. Qiu, H. Zhang, B. Yao, W. Qi, D. Jiang, W. Chen, and N. Duan. Coderetriever: A large scale contrastive pre-training method for code search. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 2898–2910, 2022. 3

[24] C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu, and T.-Y. Lin. Magic3d: High-resolution text-to-3d content creation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 300–309, 2023. doi: 10.1109/CVPR52729.2023.00037 2

[25] D. Lindlbauer, A. M. Feit, and O. Hilliges. Context-aware online adaptation of mixed reality interfaces. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, UIST '19, p. 147–160. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3332165.3347945 2

[26] R. Ma, A. G. Patil, M. Fisher, M. Li, S. Pirk, B.-S. Hua, S.-K. Yeung, X. Tong, L. Guibas, and H. Zhang. Language-driven synthesis of 3d scenes from scene databases. *ACM Trans. Graph.*, 37(6), Dec. 2018. doi: 10.1145/3272127.3275035 2

[27] Microsoft. Mixed reality toolkit, 2023. Accessed: 2025-03-12. 2

[28] Microsoft. Playwright, 2023. 4

[29] M. Mukherjee and V. J. Hellendoorn. Sosecure: Safer code generation with rag and stackoverflow discussions, 2025. 4

[30] M. R. Parvez, W. U. Ahmad, S. Chakraborty, B. Ray, and K.-W. Chang. Retrieval augmented code generation and summarization, 2021. 3

[31] N. Pereira, A. Rowe, M. W. Farb, I. Liang, E. Lu, and E. Riebling. Arena: The augmented reality edge networking architecture. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 479–488, 2021. doi: 10.1109/ISMAR52148.2021.00065 2

[32] J. C. Pinheiro and D. M. Bates. *Mixed-effects models in S and S-PLUS*. Springer, New York, NY [u.a.], 2000. 6

[33] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion, 2022. 2

[34] W. Qian, C. Gao, A. Sathya, R. Suzuki, and K. Nakagaki. Shape-it: Exploring text-to-shape-display for generative shape-changing behaviors with llms. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, UIST '24. Association for Computing Machinery, New York, NY, USA, 2024. doi: 10.1145/3654777.3676348 3

[35] J. Roberts, A. Banburski-Fahey, and J. Lanier. Surreal vr pong: Llm approach to game design. In *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, December 2022. 3

[36] J. Seo, S. Hong, W. Jang, M.-S. Kwak, H. Kim, D. Lee, and S. Kim. Retrieval-augmented text-to-3d generation, 2024. 3

[37] C. Sun, J. Han, W. Deng, X. Wang, Z. Qin, and S. Gould. 3d-gpt: Procedural 3d modeling with large language models, 2024. 3

[38] Unity Technologies. Unity, 2023. Game development platform. 2

[39] Y. Wang, S. Guo, and C. W. Tan. From code generation to software testing: Ai copilot with context-based rag. *IEEE Software*, pp. 1–9, 2025. doi: 10.1109/MS.2025.3549628 4

[40] Z. Z. Wang, A. Asai, X. V. Yu, F. F. Xu, Y. Xie, G. Neubig, and D. Fried. Coderag-bench: Can retrieval augment code generation?, 2025. 3

[41] Z. Xiao, X. He, H. Wu, B. Yu, and Y. Guo. Eda-copilot: A rag-powered intelligent assistant for eda tools. *ACM Trans. Des. Autom. Electron. Syst.*, Jan. 2025. Just Accepted. doi: 10.1145/3715326 4

[42] A. XR. arena-py: Python library for accessing the arena. `https://github.com/arenaxr/arena-py`. 4

[43] Y. Xu, Y. Ng, Y. Wang, I. Sa, Y. Duan, Y. Li, P. Ji, and H. Li. Sketch2scene: Automatic generation of interactive 3d game scenes from user's casual sketches. *arXiv preprint arXiv:2408.04567*, 2024. 2

[44] Y. Yang, F.-Y. Sun, L. Weihs, E. VanderBilt, A. Herrasti, W. Han, J. Wu, N. Haber, R. Krishna, L. Liu, C. Callison-Burch, M. Yatskar, A. Kembhavi, and C. Clark. Holodeck: Language guided generation of 3d embodied ai environments, 2024. 3

[45] Z. Yang, S. Chen, C. Gao, Z. Li, X. Hu, K. Liu, and X. Xia. An empirical study of retrieval-augmented code generation: Challenges and opportunities. *ACM Trans. Softw. Eng. Methodol.*, Feb. 2025. Just Accepted. doi: 10.1145/3717061 4, 7

[46] J. Zamfirescu-Pereira, E. Jun, M. Terry, Q. Yang, and B. Hartmann. Beyond code generation: Llm-supported exploration of the program design space. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. Association for Computing Machinery, New York, NY, USA, 2025. doi: 10.1145/3706598.3714154 3

[47] L. Zhang and S. Oney. Flowmatic: An immersive authoring tool for creating interactive scenes in virtual reality. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20, p. 342–353. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3379337.3415824 2

[48] L. Zhang, J. Pan, J. Gettig, S. Oney, and A. Guo. Vrcopilot: Authoring 3d layouts with generative ai models in vr. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, UIST '24. Association for Computing Machinery, New York, NY, USA, 2024. doi: 10.1145/3654777.3676451 3