ChatWorld: Embodied Multi-Agent Simulation For Extensible Synthesis of Hypermedia

Adrian "Avaer" Biedrzycki[*1,*], Shaw "Moon" Walters[2], and Seung Dong ("PENDINGREALITY")[3]

[2]Autonomous Research Group
[3]M3 Organization

August 28, 2023

[*]Lead author

Contents

1 Introduction **4**
- 1.1 Convergence of User and Agent Systems . . . 4
- 1.2 Simulation Systems as a Data Platform . . . 5
- 1.3 Success Metrics and Competing Forces . . . 5

2 Related Work **5**
- 2.1 Virtual Worlds . . . 5
- 2.2 Simulation Networking . . . 6
- 2.3 Multi-agent Simulation . . . 6
- 2.4 Human-AI Interaction . . . 6
- 2.5 Embodied LLM Agents . . . 7
- 2.6 Generative World . . . 7
  - 2.6.1 Rendering Interface . . . 7
  - 2.6.2 Networking System . . . 8
- 2.7 Architecture Overview . . . 8
- 2.8 World Rendering . . . 10
- 2.9 Visual Perception Loops . . . 10
- 2.10 Generative Assets . . . 11
- 2.11 2D Character Generation . . . 11
- 2.12 World Generation . . . 12
  - 2.12.1 Depth-aware 3D Skybox/Diffusion Based Virtual World Generation . . . 12
  - 2.12.2 Monocular Depth Estimation . . . 13
  - 2.12.3 Continuous Guided Video Model . . . 14
  - 2.12.4 Unguided Video Model . . . 14
- 2.13 Audio Generation . . . 14
- 2.14 Sound Effects Generation . . . 15

3 Embodied Agents **15**
- 3.1 Agent Model Interface . . . 15
- 3.2 Agent Loop . . . 15
- 3.3 Agent SDK . . . 16

4 User Interface **16**
- 4.1 User Actions . . . 16

5 System Interface **16**

6 Applications **17**
- 6.1 Companion Agent . . . 17
- 6.2 Interactive Virtual Game World . . . 17
- 6.3 "Infinite Anime" TV Show . . . 18
- 6.4 Self-Programming Agent . . . 19
- 6.5 Multi-actor Decision Agent . . . 19
  - 6.5.1 Real-world Embodiment . . . 19
  - 6.5.2 Financial Imbuement . . . 20

7 Conclusion **20**

8 Acknowledgements **20**

9 References **21**

10 Appendix **22**
10.1 Image gallery 23

Abstract

Multi-agent artificial intelligence systems powered by large language models have proven to exhibit superior performance over single-agent systems.

Our work is inspired by prior art such as Generative Agents [10], Voyager [14], and SHOW-1 [9].

Systems that hybridize AI-human interaction demonstrate impressive steerability. Such research has shifted discussion of the fusion of human intelligence and Artificial Intelligence (AI) from theoretical to prescient.

We introduce **ChatWorld**, a system for an embodied web-based virtual world "metaverse" with programmable multi-modal input/output connected to AI agents, realtime human actors and downstream computational workflows. Our system is designed as a structural template for productizable human-AI interaction.

Building upon recent advances in large language models, web technologies, and 3D game engines, our architecture enables flexible rendering and simulation through an extensible plugin and socket system. Network participants and embodied agents can engage in a shared simulation using standard web APIs, accessible through a conventional web browser.

We describe our implementation, including virtual world generation using classical procedural and generative AI techniques, multi-agent simulation, and world understanding through computer vision and other techniques. We describe the leanings and challenges of our implementation. Additional applications of ChatWorld are explored, including integration with third-party agents and portability to other systems and game engines using a sandboxed JavaScript environment.

We hope our findings and methodologies inspire further exploration and innovation among AI researchers, virtual world creators, and game developers, fostering positive collaboration between humans and rapidly advancing embodied AI technology.

Keywords

Human-AI interaction, agents, generative AI, large language models, virtual reality, open-world simulation, semantic understanding, computer vision, networked simulation, MMO video game.

1 Introduction

The integration of artificial intelligence (AI) into human-centric systems has ushered in a new epoch in interactive computing. This confluence is characterized by unprecedented interplay between human beings, AI agents, and heterogeneous programmable agent systems producing simulacra that are increasingly difficult to separate from human productivity. This is not merely a technological advancement; it represents a qualitative shift in human-machine interaction. This paper delves into the complex interconnections within this domain, focusing on the convergence of user and agent systems, the role of simulation environments.

We introduce **ChatWorld**, a massively multiplayer agent simulation framework designed to foster the positive development of this technology, and **Upstreet.ai**, a prototype implementation implmeneted as a massively multiplayer online video game.

1.1 Convergence of User and Agent Systems

In the paradigm of multi-agent systems, agents are interactive beings capable of engaging with

human users and other agents within shared virtual worlds. It brings to the fore critical questions related to ethics, privacy, and safety. How do we ensure that human-centric approaches are preserved? What are the moral responsibilities in shaping this integration? How do we allow users to practically test and iterate on these problems? These issues are not ancillary but central to the development of a responsible and human-compatible AI ecosystem[1].

1.2 Simulation Systems as a Data Platform

Simulation systems are fertile ground for generating and analyzing practical data of human-computer and increasingly agent-agent interaction. By constructing controlled virtual realities, it is possible to model, study, and optimize the diverse interactions between humans, AI agents, and holistic multi-agent systems. These simulated environments serve as both laboratories and playgrounds, where theoretical concepts can be tested qualitatively and training data can be collected.

As virtual world simulations become more realistic, research increasingly finds real-world application. Consumer-oriented mixed reality hardware such as the Apple Vision Pro can plausibly diffuse applications of this technology to a mainstream market, fostering increased investment into research and development of ancillary simulation technology.

1.3 Success Metrics and Competing Forces

Our goal function in development of ChatWorld is two-faceted: to provide an environment for practical AI agents to interact with each other, and an interpretable framework for users to interact with them.

Underpinning the development and functionality of generative AI systems like ChatWorld are three competing forces: productivity, generativity, and steerability. Productivity refers to the system's ability to perform useful work, encompassing both economic value and entertainment aspects. Generativity, characterized by unforeseen behaviors, distinguishes AI from traditional heuristic-based models, while introducing unique challenges[2]. Steerability involves the balance of aligning user intentions with system design, often necessitating substantial engineering efforts[3].

The exploration of these domains forms the core of our research. Through rigorous analysis, innovative methodologies, and the creation of practical applications, we seek to contribute to the understanding of the complex dynamics at play in the rapidly evolving field of human-agent collaboration. The following sections of this paper will delve into specific aspects, methodologies, and findings, illuminating the path forward in this exciting arena of study.

2 Related Work

This section outlines the existing work and foundational concepts that underpin our innovative approach in ChatWorld. It provides an overview of virtual worlds, world simulation networking, multi-agent simulation, human-AI interaction, and the integration of large language models with embodied agents.

2.1 Virtual Worlds

Virtual worlds like as VRChat, Roblox, Decentraland, and Horizon provide sensory-immersive environments that allow humans to interface with a networked simulation. AI agents are increasingly taking part in these worlds. Even if the simulation does not provide support for doing so, AI agents are able to use technologies like modding tools to add LLMs and text to The Elder Scrolls 5: Skyrim [4], use OSC to create virtual reality agent experiences in VRChat [13], and even use raw video capture to simulate a

[1] See, e.g., Russell, S., et al., "Human Compatible: Artificial Intelligence and the Problem of Control," (2019)

[2] For an exploration of generativity in AI, see: Goodfellow, I., et al., "Generative Adversarial Nets," arXiv preprint arXiv:1406.2661 (2014)

[3] See, e.g., OpenAI's work on AI alignment: "Alignment Research: A Long-Term Strategy," OpenAI Blog (2021)

sensory environment for the agent to embody for competitive purposes, as in video game bots [4].

While not necessarily inhabiting the same space as players, virtual worlds provide an embodiment interface for humans to form productive empathic relationships with artificial intelligence agents that mimic some physical human behaviors. By giving agents roles and personalities that can be configured, users can exert agency over the agents to configure them to their preference.

This model has already seen significant usage in low-fidelity products like character.ai [2], but it is likely that increasing fidelity and interoperability in embodied virtual worlds will expand the domain of application of such agents.

2.2 Simulation Networking

The architecture of networked multiplayer systems, particularly within the realm of multi-agent systems and MMO games, is pivotal to the development of ChatWorld. Networked simulation serves as the backbone for the coordination and seamless integration of agents and human participants within a shared virtual environment. The drive to create immersive and real-time interactive environments necessitates robust and scalable network architectures. This transition from isolated agent simulations to networked multi-agent environments enables more complex interactions and realism[5].

Challenges such as latency, synchronization, and security are paramount. Balancing efficiency and responsiveness without compromising the integrity of the simulation is an ongoing challenge. Historical efforts in MMO games and networked simulations, such as EVE Online and World of Warcraft, provide insights into handling large-scale, distributed systems. These experiences guide the design of the ChatWorld networking framework, making it capable of supporting intricate interactions between multiple players and agents[6].

2.3 Multi-agent Simulation

Multi-agent systems (MAS) are central to understanding and exploring complex interactions within a shared environment. The application of MAS in ChatWorld facilitates nuanced interactions between diverse agents. Simulating multiple intelligent agents allows the modeling of complex systems and scenarios that mirror real-world interactions and behaviors. MAS has been utilized in various domains such as economics, social sciences, and robotics, thus forming a rich foundation for our exploration[7].

Coordination, communication, and learning among agents present intricate problems, requiring careful design and implementation. Research on multi-agent reinforcement learning and social dilemmas provides a rich background, and the integration of language models with MAS adds a new layer of complexity and opportunity. The use of natural language understanding within multi-agent systems opens up new avenues for collaboration and decision-making, furthering the scope of what can be achieved within virtual worlds[8].

2.4 Human-AI Interaction

The interplay between human users and AI agents is an evolving field, exploring the dynamics, ethics, and practicalities of human-machine collaboration. Enhancing the synergy between human and AI agents promotes more effective collaboration, empathy, and usability. This collaboration aims to go beyond mere interaction,

[4]Example of a video game bot using computer vision: https://github.com/RootKit-Org/AI-Aimbot

[5]Bernier, Y., "Latency Compensating Methods in Client/Server In-game Protocol Design and Optimization," Game Developers Conference (2001)

[6]Bryant, D. and Iizuka, H., "Player Behavior and Traffic Characterization for MMORPGs," Network and Systems Support for Games (2005)

[7]Wooldridge, M., "An Introduction to MultiAgent Systems," John Wiley and Sons (2009)

[8]Leibo, J. Z., et al., "Multi-Agent Reinforcement Learning in Sequential Social Dilemmas," arXiv preprint arXiv:1702.03037 (2017)

[9]Cassell, J., et al., "Embodied Conversational Agents," MIT Press (2000)

forming partnerships where both human and AI agents can achieve shared goals[9].

Building trust, ensuring ethical interactions, and designing intuitive interfaces are essential yet complex tasks. Research in human-computer interaction, cognitive psychology, and AI ethics informs the design principles for ChatWorld, promoting user-friendly and responsible agent behaviors. AI Dungeon and other interactive narrative systems demonstrate the possibilities of engaging human-AI collaboration, forming a basis for further exploration within the realm of virtual worlds[10].

2.5 Embodied LLM Agents

The integration of large language models with virtual embodiment provides a novel paradigm for agent behavior and interaction. Coupling large-scale language models with embodied agents enriches the expressiveness and understanding of AI systems within virtual environments, bridging the gap between text and spatial representation. This integration enables more natural and human-like interactions, opening up new possibilities for research and application in various fields, such as education, entertainment, and virtual assistance[11].

Aligning text-based understanding with visual and spatial comprehension requires intricate modeling and engineering. The challenge of creating agents that can comprehend and navigate both textual and visual information is substantial, yet it provides a pathway to more immersive and intelligent systems. PaLM-E, Voyager, and other projects have pioneered the integration of text-based AI models with 3D environments. These innovations form the cornerstone for the design of ChatWorld's agents, facilitating a rich and engaging user experience[12].

2.6 Generative World

ChatWorld's generative world is built on popular web technologies (Three.js, Node.js) and runs within a browser environment. The simulation layer mirrors common MMO paradigms and focuses on 3D embodiment, interoperability, social interaction, multimodal simulation, and agent-world interaction. Its design supports asset translation across platforms and domains, including real-world embodiment through APIs. This architecture promotes a seamless integration of diverse media, enhancing the realism and engagement of the virtual world[13].

Challenges in bridging the gap between traditional text-based descriptions and 3D virtual environments, while maintaining scalability and responsiveness, are complex. The integration of text descriptions in MUDs and the application of OCR and CV techniques for scene capture have inspired the design of ChatWorld's simulation layer. These advancements pave the way for a more dynamic and responsive virtual world, where agents can adapt and respond to a constantly changing environment. By embracing these challenges, ChatWorld pushes the boundaries of virtual collaboration and immersive computing, offering a glimpse into the future of human-agent interaction within shared virtual spaces.

2.6.1 Rendering Interface

ChatWorld's rendering interface is a crucial component that translates the simulation state into an audio-visual stream. This stream can be displayed to the user in real-time or rendered for offline use. The rendering interface utilizes a browser-based system that can run either in headful or headless mode, catering to various user preferences and system requirements.

The interface is responsible for displaying 3D representations of agents and the environment, GUI overlays for controlling the simulation, and

[10]Walton, R., "AI Dungeon: A Novel Approach to Text-Based Adventure," Latitude (2019)

[11]Shawar, B. A., and Atwell, E., "Chatbots: Are they Really Useful?," LDV Forum (2007)

[12]Su, W., et al., "Voyager: Exploratory Analysis via Faceted Browsing of Visualization Recommendations," IEEE Transactions on Visualization and Computer Graphics (2017)

[13]Bartle, R., "Designing Virtual Worlds," New Riders (2003)

rendering real-time speech and audio effects. A noteworthy feature of the ChatWorld rendering interface is the integration of an LLM-controlled camera system. This system enables automated focus when presenting the output of the simulation, allowing for cinematic perspectives and enhancing the visual immersion of the environment. Such an approach is aligned with recent trends in virtual world rendering, providing both aesthetic and functional benefits[14].

2.6.2 Networking System

ChatWorld's networking system is an exemplar of innovative design, providing robust support for multiple users to connect to the simulation and interact with agents. Leveraging a spatially-mapped cluster of CloudFlare Workers, the architecture is scalable and can accommodate an arbitrary number of users distributed throughout the game world.

The system draws on established concepts from Conflict-free Replicated Data Types (CRDTs) and Peer-to-Peer (P2P) networking systems, ensuring smooth client-side interpolation and replication for all connected clients[15]. This approach facilitates real-time multiplayer position handling and robust communication via text and voice chat interfaces.

One of the innovations in ChatWorld is the concept of "stitched multiplayer realms." Spatial handoff across the world space is realized through a sophisticated replication and deduplication of user data across server realms, guided by spatial coordinates. This ensures seamless transitions and contributes to the illusion of a continuous virtual space.

The architecture also includes networked physics handling and ownership mechanisms that ensure the eventual consistency of the simulation. This includes an automatic ownership-based model that resolves conflicts and maintains the integrity of the virtual world. While there are inherent limitations and challenges, such as the potential for hacking within P2P CRDT systems, ongoing research and development are aimed at identifying and implementing effective mitigation.

2.7 Architecture Overview

The architecture of ChatWorld is constructed around a tickable agent model game loop, a design that allows both user inputs and agent actions to dynamically influence the progression of the simulation.

Actions and memories are meticulously logged for each agent within a memory database which supports similarity search and retrieval with vectors [6]. This architecture provides a rich history that can be queried by registered agent models, facilitating the synthesis of prompts and the production of agent actions. This logging system serves as a foundational component that supports both the immediate interactions within the virtual environment and the long-term evolution of agent behaviors.

ChatWorld's user interface integrates several input devices and modalities including screen rendering, keyboard, mouse, camera and microphone inputs. These inputs are translated into triggers for the agent model, providing an intuitive and responsive interface for users to engage with the virtual world.

Complementing the user interface, the system interface manages state for the world, items, characters, and lore. These components collectively feed into the context of the agent model, enabling a rich and nuanced interaction between agents and their environment. Conversations, as a critical aspect of human-agent interaction, are stored and accessible within a memory database, allowing agents to draw on past interactions and build more coherent and engaging narratives.

The ChatWorld Agent SDK serves as a bridge, facilitating input and control between the agent code and the IO controller. This ensures seamless integration between various components of the system, enhancing the overall coherence and functionality of the platform.

[14] Andujar, C., et al., "Automatic Camera Control in Virtual Environments," Computers and Graphics (2002)

[15] Shapiro, M., et al., "A comprehensive study of Convergent and Commutative Replicated Data Types," INRIA (2011)

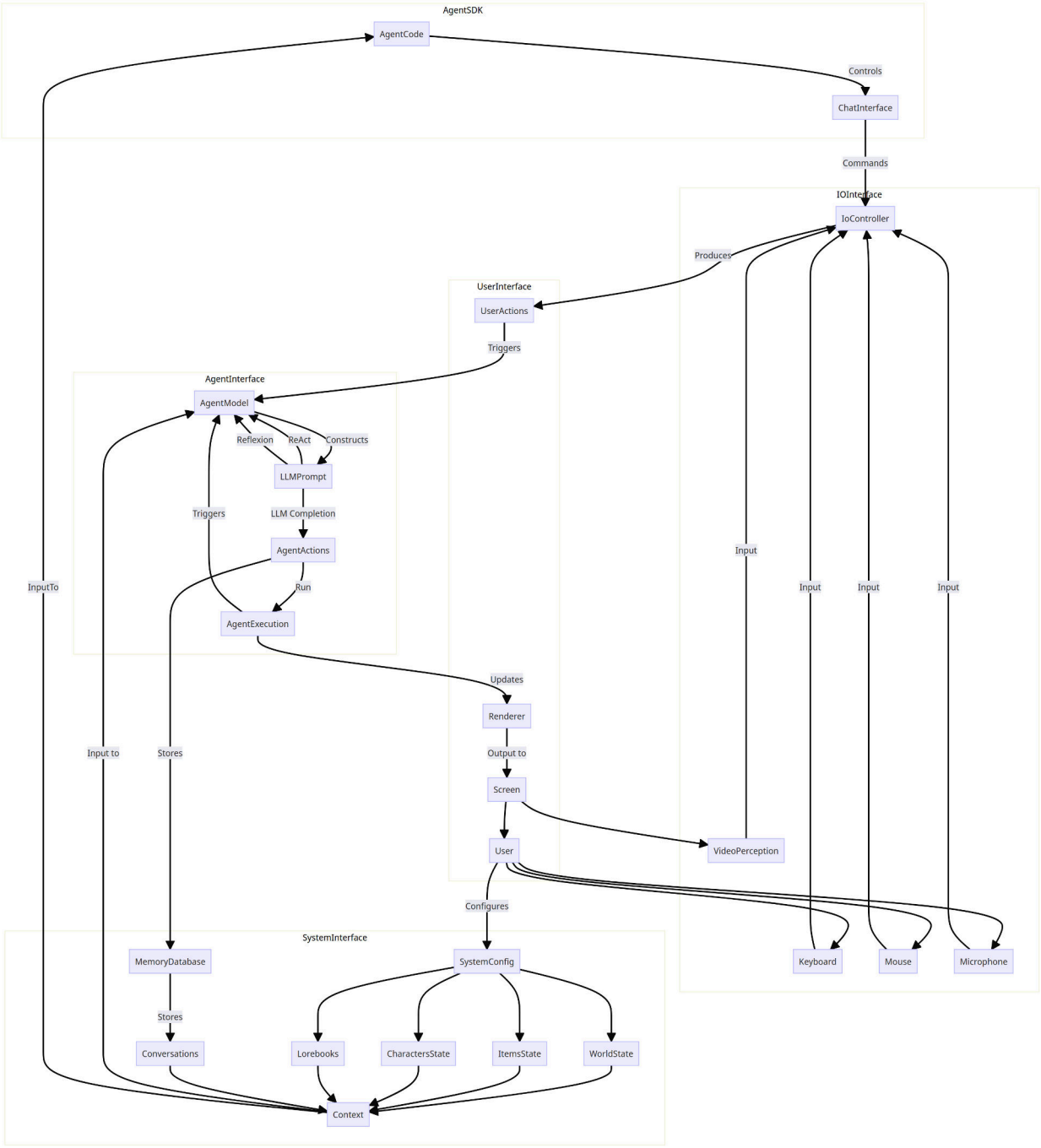

Figure 1: Architecture diagram of the ChatWorld system

Animation, expression, locomotion and interaction are all processed and visualized by the core engine and rendered in realtime at a minimum of 30 frames per second. All events are propagated to other agents through the multiplayer system, which is a custom CRDT store running on lightweight Cloudflare Durable Objects. The design enables a virtually limitless number of agents, provided they are spread out to an approximate density of 50 agents per $1/2\text{km}\hat{2}$.

2.8 World Rendering

World rendering in ChatWorld encompasses the methods employed for visually representing the virtual environment. The rendering process is responsible for translating the underlying simulation state into a 3D representation that includes agents, objects, landscapes, and various graphical overlays. The system's ability to run both headful and headless provides versatility in rendering modes, catering to different use cases and computational resources. Real-time speech and audio effects further contribute to an immersive experience, blending visual aesthetics with auditory engagement.

2.9 Visual Perception Loops

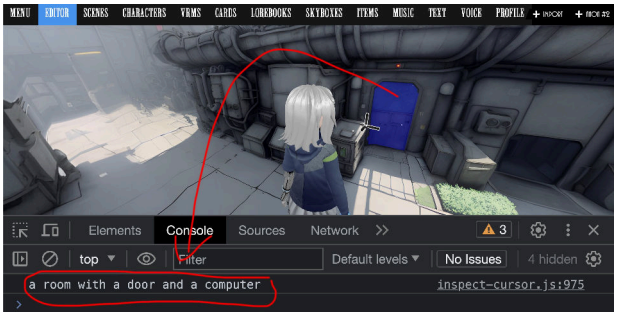

Figure 2: Object recognition in Diffusion Based Virtual World Generation

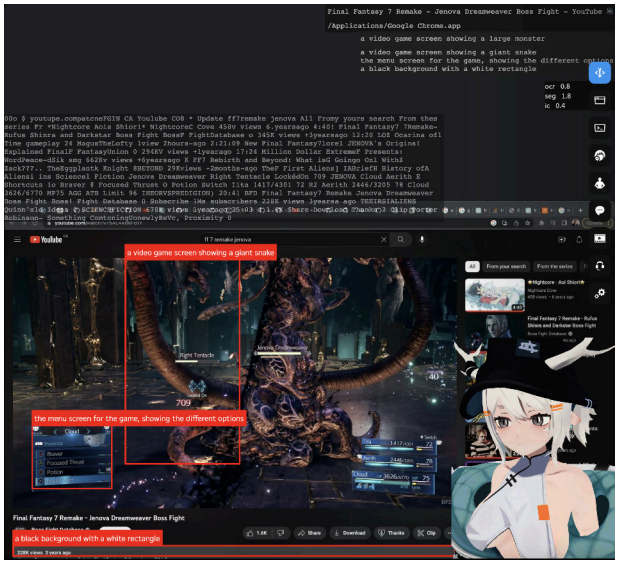

Figure 3: Real time OCR using browser stream as visual input

The visual perception loops in ChatWorld serve as a bridge between the rendering process and the AI agents. These loops provide mechanisms for AI agents to perceive and interpret the visual elements within the virtual environment. Through computer vision and optical character recognition (OCR) techniques, agents can capture and understand their environment from both an 'in-world' perspective as well as an 'on-screen' perspective. For example, the world system provides the agent with world data in world space relative to them, while the OCR can process text and regions in 2D space to give the agent awareness of the world from the user interface perspective. OCR is applied to the whole screen to extract and identify visual features, objects, and text.

The integration of visual perception loops with language models adds a new dimension to agent understanding, enabling more nuanced and context-aware interactions. The scope of visual context is not limited to the simulated world, but any video stream (such as from a user's desktop, phone camera, or webcam). The alignment of text-based comprehension with vi-

[16]Zhu, Y., et al., "Visual Semantic Role Labeling: A New Task, Analysis and Model," arXiv preprint arXiv:1605.02001 (2016)

sual perception allows for human-like agent behaviors and enriches the research possibilities within virtual worlds[16].

2.10 Generative Assets

Figure 4: 2D generated sprite sheet representation of a tool

Generative assets occupy a unique intersection of aesthetics and functionality within the architectural framework of ChatWorld. Although not a part of the core simulation loop, multi-modal generative models are employed to create a diverse array of game assets, encapsulating both visual and auditory elements.

The generative AI pipeline offers dynamic rich visual content to enhance the immersive experience. Practically, generated assets are processed through a perception layer that provides the virtual characters with a simulated sensory perception, or "eyes in the world." This includes semantic and metadata information about the asset.

The generation of in-game items and characters follow similar methods. However, items necessitate specific fine-tuning. By adapting the noise diffusion technique used for character synthesis, we have developed a pattern particularly suited for item image generation. The customization of diffusion processes for various elements within the virtual environment illustrates the potential for nuanced control and specialization in content generation[17].

While the full utilization of this aspect remains a subject for future exploration, the underlying methodology has shaped the current system's design, offering a fertile ground for innovation in human-machine interaction and virtual embodiment.

2.11 2D Character Generation

Figure 5: Pre-denoise representation for character sheet

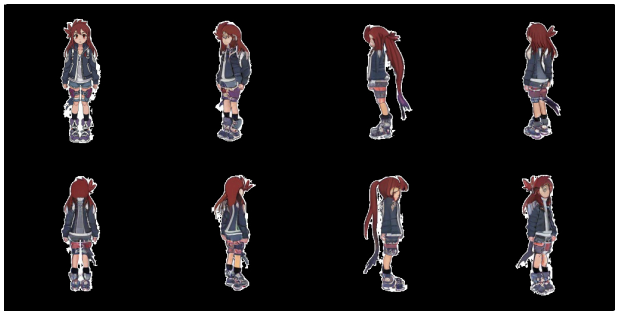

Figure 6: 360 Generated Character Spritesheet

Figure 7: Generated Character Emotion Sprites

[17]Kingma, D.P., et al., "Variational Inference with Normalizing Flows," ICML (2015)

Character generation in ChatWorld is a nuanced process that combines artistry with machine learning. Utilizing Stable Diffusion fine-tuned image generation models seeded via noise-based mask vectors, high-quality artistic results are achieved. Separate masks for male and female characters ensure consistency in the generated images.

The generation process extends to synthesizing dynamic facial expressions, including mouth flaps and affective cues. Utilizing a specialized background removal AI model, characters are precisely extracted for in-world presentation.

The resulting AI model yields prompted character synthesis with an impressive acceptance rate exceeding 90

The application of the Zero-1-to-3[8] image reprojection model generates multiple character viewpoints, adapting to camera angles within a 3D sprite-based system. Future implementations may leverage models like ControlNet to enrich character poses, with potential extensions to more intricate 3D models.

2.12 World Generation

The believability of agent systems depends on virtual embodiment, in which characters demonstrate a coherent understanding of their world, as humans do. In ChatWorld, virtual characters navigate using actions derived from Large Language Models (LLMs). This requires a congruent description of the environment that aligns with human users' perception.

A model based on succinct descriptions of in-world assets is employed, allowing creative flexibility while maintaining a unified virtual experience. For instance, agents may describe the trees in the world in stylized terms such as "anime-style."

This approach lays the groundwork for future innovations in world generation. The theoretical capacity to synthesize limitless settings for simulation extends the potential for agents to explore an infinite "dreamworld" canvas.

The alignment of generative algorithms with narrative context marks a pioneering step in virtual simulations. The techniques and insights gleaned from ChatWorld character and world generation offer a glimpse into the tremendous potential of AI in crafting immersive virtual realities.

2.12.1 Depth-aware 3D Skybox/Diffusion Based Virtual World Generation

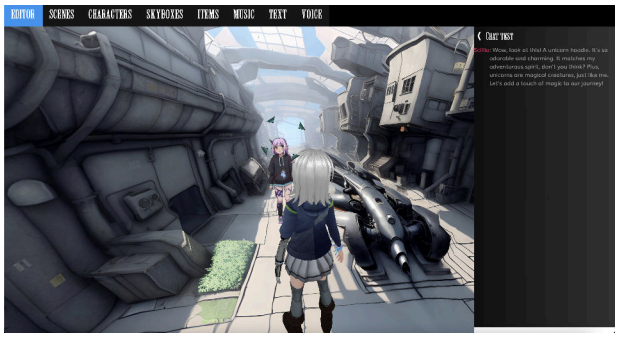

Figure 8: Diffusion Based Virtual World Generation 1

Figure 9: Diffusion Based Virtual World Generation 2

Figure 10: Diffusion Based Virtual World Generation 3

2.12.2 Monocular Depth Estimation

Figure 11: Monocular Depth Estimation 1

Figure 12: Monocular Depth Estimation 2

Blockade Labs [18] introduces an innovative prompt-based model for synthesizing a complete 3D scene, providing an elegant solution to integrating depth into the virtual environment. The depth-projected skybox, although rudimentary in nature, serves as a foundational layer for world physics.

The fidelity of the generated art is not only visually appealing but offers functional benefits. Agents are endowed with spatial awareness within the skybox and utilize BLIP-2[7] in the perception stack to facilitate navigation to specific points of interest.

An intriguing feature of this model is the ease of adding seamless portals to the generated skyboxes. This capability paves the way for the creation of a continuous, interconnected world for agents, enhancing the dynamism and complexity of the simulation.

The integration of depth AI models, such as MiDaS[11] with Stable Diffusion, enables the

[18] https://skybox.blockadelabs.com/

transformation of arbitrary images into a virtual world projected in a comic-style 3D space. This process, although appearing complex, is executed through a straightforward depth deformation of an image guided by the depth map.

The resulting scene unlocks the possibility for 3D characters to traverse an arbitrary 2D space, with applications ranging from real-world photography to hand-drawn comic panels.

The implementation of this approach posed several challenges:

1. **Ground Navmesh Detection**: Achieved through floor/ground segmentation and normal vector alignment.

2. **Camera Intrinsics Computation**: Utilizing AI trained on camera intrinsic data ensured accurate projection of the 3D scene.

3. **Scene Bounding**: Ensuring characters remain within the panel required physics wall planes that conform to the scene's bounds.

4. **Occlusion Physics Filling**: To avoid jarring occlusion physics, a strategy involving depth discontinuities detection and top-down perspective render was employed.

5. **Scene Scaling**: The non-scale invariant nature of the MiDaS depth model posed a challenge for scene scaling, requiring manual user intervention. Another model, ZoeDepth[1], although scale-invariant, did not readily facilitate scene scaling without ground truth reference.

We implement "3D inpainting", a process of continuous erasure of regions of color and depth and infilling of new color and depth information. We generate new view-centric depth projections to produce a physical mesh of the environment which agents can interact with via ChatWorld's physics engine. To ensure seamless integration, depth-aware generative models are employed to facilitate the creation of coherent new spaces. This continuous approach to 3D inpainting not only amplifies the spatial dimension but also reinforces the authenticity and richness of the depth cues in the augmented regions.

Together, the Skybox Depth Model and the Comic Panel Depth Model contribute to the dynamic environment within ChatWorld. By leveraging cutting-edge technologies and innovative approaches, these models enhance the realism and interactivity of the virtual space, providing a rich backdrop for agent interactions and exploration.

2.12.3 Continuous Guided Video Model

The exploration of procedural scene generation has led to the integration of a continuous guided video model, such as Deforum[19]. This model leverages the principles of Stable Diffusion to synthesize arbitrary video scenes through a two-step process: a guided camera pass for video output, followed by a depth pass. While the resulting scenes may lack complete stability, they manifest unique artistic qualities. Depending on the application's requirements and the desired aesthetic, this technique offers a viable option.

2.12.4 Unguided Video Model

AnimateDiff, a contrasting model, functions without guidance to generate arbitrary short video clips from a single prompt[20]. Though unguided, its ability to produce aesthetically pleasing results has proven valuable in various artistic contexts. The exploration of unguided models like AnimateDiff adds another dimension to the video generation techniques, allowing for more flexible and spontaneous content creation.

2.13 Audio Generation

The auditory dimension of the virtual environment is enriched through the implementation of Meta AI's MusicGen [3]. With the capability to produce creative audio syntheses from a single text prompt, this model demonstrates

[19] `https://deforum.github.io/`

[20] `https://animatediff.github.io/`

a promising direction for AI-generated audio soundtracks. Although not designed for real-time operation, the model's extensibility for producing audio of varying lengths holds potential for future advancements in scene-complementing soundtracks.

2.14 Sound Effects Generation

Complementing the music generation, we use AudioGen[5] for the production of sound effects[21]. AudioGen has shown promising results in its ability to synthesize unique sound effects, enhancing the auditory experience while significantly reducing the cost associated with building a sound effects library for games and simulations.

The combination of these various generative models contributes to a holistic and immersive virtual experience. From video synthesis to sound effect creation, the application of AI techniques has broadened the scope and intricacy of content generation within ChatWorld. The continuous exploration and integration of these models are indicative of the evolving nature of AI in shaping rich and dynamic virtual environments.

3 Embodied Agents

Figure 13: ChatWorld Agents in Upstreet

Embodied agents within ChatWorld are categorized into two kinds: internal and externally simulated. Interally simulated agents implement ReAct[15] and Reflexion[12] subsystems within the agent model. These agents are all simulated by the same system, called the Director, and are akin to traditional Non-Player Characters (NPCs) in video games.

Network-connected external entities run external to the system and connect to the game world through the network. These agents have full access to the game world data but have to carry their own memory and conversation processing subsystems. Implementations in Python and Javascript for connecting external agents can be found in the Upstreet SDK[22].

3.1 Agent Model Interface

This interface is the operational core of the agent's behavior, encompassing definition, execution, and processing. Key components include:

- **Context:** A compendium of information guiding the agent's decision-making process, including world state, conversations, and more.
- **AgentModel:** Responsible for processing context, constructing LLMPrompt, and orchestrating interactions with ReAct and Reflexion systems.
- **AgentExecution:** The execution engine, triggering the agent model and updating the Renderer to mirror changes in the virtual world.
- **AgentActions:** Resultant actions executed by agents, stored in the Memory-Database for future reference, and underpinning the AgentExecution cycle.

3.2 Agent Loop

Inspired by recent advancements in Generative Agents[23], and Voyager[24], we provide a default

[21] `https://felixkreuk.github.io/audiogen/`
[22] https://github.com/m3-org/upstreet-sdk
[23] C. Park, et al., "Generative Agents," `https://arxiv.org/abs/2304.03442`
[24] Wang, L., et al., "Voyager: An Open-Ended Embodied Agent with Large Language Models," `https://arxiv.org/abs/2305.16291`

implementation of a multi-agent simulation loop within ChatWorld. This implementation takes cues from the aforementioned works, synthesizing concepts and methodologies to create a robust framework for agent interaction, decision-making, and world manipulation. The loop facilitates the continuous evolution of agents, allowing them to learn, adapt, and grow within the confines of the simulated world.

3.3 Agent SDK

The Agent Software Development Kit (SDK) plays a pivotal role in enabling agent models written in JavaScript to execute and influence the virtual world. By incorporating strategies and insights from previous research, our approach demonstrates the flexibility and scalability of AI-model-agnostic-agent integration within ChatWorld.

For example, the methodologies outlined in the "Generative Agents" research paper, including its multi-step prompt loop, can be seamlessly implemented and loaded as an agent model in ChatWorld. The platform's architecture supports the parallel execution of multiple agent models, fostering an environment of collaboration, experimentation, and innovation. Such parallelism allows for the simultaneous mixing and re-mixing of implementations, reducing the time needed to test, debug, and visualize research.

Furthermore, ChatWorld provides APIs for agent memory, perception, events, and agent action triggers. Coupled with a sandboxed JavaScript execution environment, this design allows for a wide array of oracle (external API) and timing strategies to be implemented and explored. The fusion of these elements creates a rich toolset for researchers, developers, and enthusiasts to craft intelligent, responsive, and complex agents within the ChatWorld ecosystem.

By combining the principles of embodiment, memory, adaptability, and modularity, ChatWorld's agent framework sets the stage for a new era of virtual interaction and simulation. The convergence of these various components underscores the potential and versatility of AI-driven virtual environments, opening doors to uncharted territories of exploration, creativity, and innovation.

4 User Interface

Figure 14: UI System in Upstreet.ai

The user interface (UI) in ChatWorld is an intricate assembly of input devices including the Keyboard, Mouse, and Microphone, all integrated through an Input/Output (I/O) Controller. This subsystem is pivotal in translating human interactions into digital signals, thereby bridging the gap between users and the virtual environment.

4.1 User Actions

User actions are digital signals generated by the I/O Controller in response to physical inputs from the user. These actions are then channeled into the agent model, where they act as triggers for specific behaviors. The seamless translation of human gestures into agent responses epitomizes the interactive nature of ChatWorld, fostering a symbiotic relationship between the user and the virtual agents.

5 System Interface

The System Interface is a cornerstone of the ChatWorld architecture, providing the necessary context for agent models and defining the state

and environment of the virtual world. Key components include:

- **WorldState, ItemsState, CharactersState, Lorebooks:** These encapsulate the entire state of the world, items, characters, and lore, forming the foundation of the virtual universe.
- **MemoryDatabase:** Central to the personalized experience within ChatWorld is the sophisticated memory subsystem. This mechanism is responsible for storing and retrieving information, allowing for continuity in agent behavior and world state. By maintaining historical data and contextual insights, the memory subsystem enables agents to recall past interactions, recognize patterns, and adapt to individual user preferences. The result is a tailored and coherent experience, where actions and decisions have lasting consequences and contribute to the unfolding virtual reality.
- **Conversations:** Integral to the ongoing narrative, conversations are woven into the fabric of the simulation, enhancing the realism and continuity of user-agent interactions.

6 Applications

ChatWorld allows for practical applications for real time, multi modal agents across various domains including entertainment, research, education, and more. The following subsections delve into some of these applications, illuminating the broad spectrum of possibilities that ChatWorld offers.

6.1 Companion Agent

The Companion Agent application within ChatWorld offers a personalized and empathetic AI-driven companion that can interact, learn, and grow with individual users. This application could be instrumental in various scenarios, such as providing emotional support, assisting with daily tasks, or serving as an educational mentor. The underlying AI models ensure a human-like interaction, adapting to user preferences and needs over time. The potential of Companion Agents to mitigate feelings of loneliness or to provide consistent support showcases a novel and humane application of artificial intelligence.

6.2 Interactive Virtual Game World

Figure 15: Upstreet.ai Story 1

Figure 16: Upstreet.ai Story 2

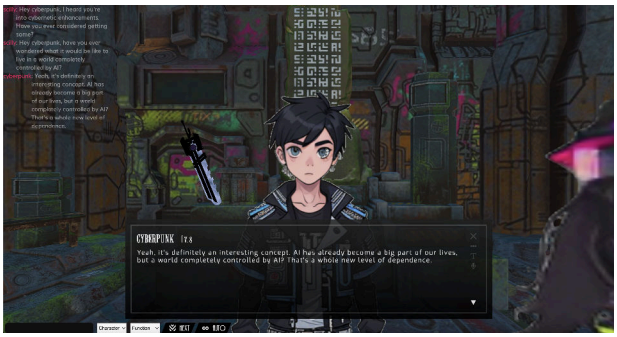

Figure 17: Upstreet.ai Story 3

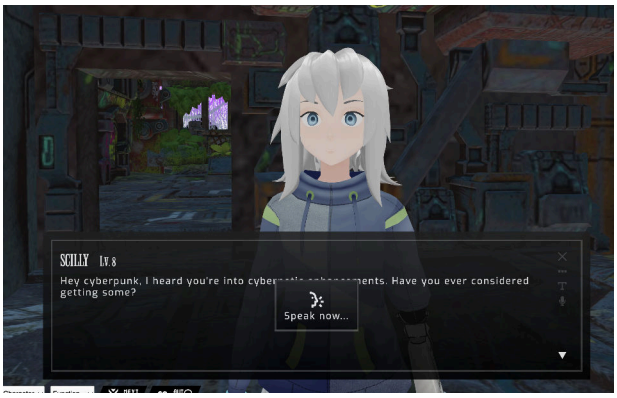

Figure 18: Upstreet.ai Story 4

Upstreet.ai, the application of ChatWorld as a Massively Multiplayer Online (MMO) video game concept presents a unique proposition for the future of virtual gaming. The current prototype validates this concept, yet the untapped potential lies in the creation of an infinitely generated virtual world, inhabited by AI-driven agents and allows the creation of a rich and dynamic environment where players can engage with intelligent Non-Playable Characters (NPCs) controlled by advanced AI models[25]. Such an environment could function indefinitely without a player base, given that the agents could be programmed to generate content and evolve the game autonomously. This concept transcends traditional gaming paradigms and offers a self-sustaining and ever-evolving virtual ecosystem. The integration of AI in this context not only enhances player engagement but also contributes to the longevity and dynamism of the virtual world.

6.3 "Infinite Anime" TV Show

Figure 19: Testing Personality Dynamics with Plot Director

The rendering capabilities of ChatWorld can also be harnessed to produce animated sequences for an "infinite" anime-style TV show. This approach utilizes generative AI models to create continuous and ever-evolving content. While there is precedent for generative AI in entertainment, the multi-agent context provided by ChatWorld offers a novel dimension. The potential to produce endless, AI-driven animated content can revolutionize the entertainment industry, offering personalized and adaptive viewing experiences.

[25]Yannakakis, G. N., and Togelius, J., "Artificial Intelligence and Games," Springer (2018)

6.4 Self-Programming Agent

Figure 20: AI Citrine, an AI Vtuber that performs coding on behalf of Twitch chat

The concept of an Self-Programming Agent within ChatWorld explores the frontier of autonomous programming. By integrating AGI models, this application can develop, test, and deploy software solutions autonomously. This extends the boundaries of traditional software development, introducing a new paradigm where human oversight is complemented by AI-driven development cycles.

Citrine is an example of an Autonomous Programming Agent within ChatWorld. It embodies a higher degree of autonomy, capable of not only coding but also understanding and interpreting complex programming paradigms. Citrine's capabilities extend to debugging, optimization, and even the creative design of new algorithms. Its integration within ChatWorld demonstrates a practical and innovative application of AI in software engineering.

6.5 Multi-actor Decision Agent

Figure 21: The Council

The Multi-actor Decision Agent, or "Jedi Council," represents an exploration of complex multi-agent decision-making processes within ChatWorld. By simulating multiple intelligent agents, each with unique perspectives and expertise, complex decisions can be analyzed and evaluated from various angles. This approach can be applied in various fields, such as business strategy, political analysis, or scientific research, offering a rich decision-making tool.

In conclusion, the diverse applications of ChatWorld reflect its versatility and potential to redefine various fields. From entertainment and gaming to autonomous programming and complex decision-making, ChatWorld stands as a testament to the innovative possibilities of multi-agent simulations and AI-driven interactions.

Examples of each of these applications will be released with the source code of ChatWorld.

In conclusion, the diverse applications of ChatWorld reflect its versatility and potential to redefine various fields. From entertainment and gaming to autonomous programming and complex decision-making, ChatWorld stands as a testament to the innovative possibilities of multi-agent simulations and AI-driven interactions.

Examples of each of these applications will be released with the source code of ChatWorld.

6.5.1 Real-world Embodiment

Beyond the virtual realm, the agent system proposed in ChatWorld can find application in real-

world embodiment such as augmented reality spaces (e.g., Vision Pro) or robotic contexts (e.g., Tesla Optimus bot). Though the mechanics of implementing such agents in the real world are beyond the scope of this paper, the techniques delineated herein offer a blueprint for the development of perceptive agents that provide intelligent services. Such agents could serve as guides in augmented reality tours, personal assistants in smart homes, or even cooperative robots in industrial settings[26]. The fusion of virtual and real-world embodiment can redefine human interaction with technology, creating seamless and personalized experiences.

6.5.2 Financial Imbuement

The intersection of ChatWorld with the financial domain opens up intriguing possibilities with profound economic implications. Envisioning agents imbued with cryptographically-signed financial assets like Non-Fungible Tokens (NFTs), one could design systems where agents trade autonomously or on behalf of their owners. This concept introduces a new dimension to algorithmic trading and financial automation[27]. Moreover, if these agents were empowered to purchase computing resources using their economic gains, they could bootstrap their economic growth, potentially emerging as significant economic entities. This scenario challenges conventional economic models and raises ethical, legal, and technical questions that warrant comprehensive exploration.

The future directions delineated above represent just a glimpse of the potential that ChatWorld holds. The fusion of gaming, real-world embodiment, and financial integration with AI-driven agents offers a fertile ground for innovation, research, and development. The exploration of these avenues could lead to novel applications and contribute to the advancement of AI, virtual reality, and economics. The flexibility and modularity of ChatWorld make it a versatile platform, poised to redefine multiple facets of human interaction with technology.

7 Conclusion

The research and development presented in this paper elucidate a novel intersection of generative AI, multi-agent simulations, and virtual reality, embodied within ChatWorld. By leveraging state-of-the-art techniques, we have opened doors to infinite possibilities within gaming, entertainment, real-world applications, and economic integration. ChatWorld's promise lies not only in its immediate applications but also in its potential to inspire future innovations, fueling both academic exploration and practical implementations.

The collaborative nature of the research community and the open-ended possibility space that ChatWorld offers make it an attractive platform for both the generative AI and game developer communities. We anticipate that the techniques and concepts presented here will ignite further interest and experimentation, leading to transformative advancements in the field.

8 Acknowledgements

Special thanks to M3 for research funding and support and contributors to Autonomous Research Group for feedback on this paper and technical work.

[26] Milani Alfredo, D., et al., "Virtual Reality and Augmented Reality in Industry 4.0," Springer (2019)

[27] Treleaven, P., et al., "Algorithmic Trading Review," Communications of the ACM, 56(11), 76-85 (2013)

9 References

References

[1] Shariq Farooq Bhat et al. *ZoeDepth: Zero-shot Transfer by Combining Relative and Metric Depth*. 2023. arXiv: 2302.12288 [cs.CV].

[2] *CharacterAI_NSFW*. Reddit community dedicated to finding ways to uncensor the chatbots on the Character.ai platform. 2023. URL: https://www.reddit.com/r/CharacterAi_NSFW/.

[3] Jade Copet et al. *Simple and Controllable Music Generation*. 2023. arXiv: 2306.05284 [cs.SD].

[4] ESO. *This AI Mod just transformed Skyrim forever!* https://www.youtube.com/watch?v=AQq8M88s3BU. Accessed: yyyy-mm-dd.

[5] Felix Kreuk et al. *AudioGen: Textually Guided Audio Generation*. 2023. arXiv: 2209.15352 [cs.SD].

[6] Patrick Lewis et al. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. 2021. arXiv: 2005.11401 [cs.CL].

[7] Junnan Li et al. *BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models*. 2023. arXiv: 2301.12597 [cs.CV].

[8] Ruoshi Liu et al. "Zero-1-to-3: Zero-shot One Image to 3D Object". In: *arXiv preprint arXiv:2303.11328* (2023). URL: https://arxiv.org/pdf/2303.11328.pdf.

[9] Philipp Maas et al. *To Infinity and Beyond: SHOW-1 and Showrunner Agents in Multi-Agent Simulations*. 2023. URL: https://fablestudio.github.io/showrunner-agents/.

[10] Joon Sung Park et al. "Generative Agents: Interactive Simulacra of Human Behavior". In: *arXiv preprint arXiv:2304.03442* (2023). arXiv:2304.03442v2 [cs.HC]. DOI: 10.48550/arXiv.2304.03442. arXiv: 2304.03442 [cs.HC]. URL: https://arxiv.org/abs/2304.03442.

[11] René Ranftl et al. *Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer*. 2020. arXiv: 1907.01341 [cs.CV].

[12] Noah Shinn et al. "Reflexion: Language Agents with Verbal Reinforcement Learning". In: *arXiv preprint arXiv:2303.11366* (2023). URL: https://arxiv.org/pdf/2303.11366.pdf.

[13] VRChat. *OSC Overview*. https://docs.vrchat.com/docs/osc-overview. Accessed: 2023-08-18. 2023.

[14] Guanzhi Wang et al. "Voyager: An Open-Ended Embodied Agent with Large Language Models". In: *arXiv preprint arXiv:2305.16291* (2023). arXiv:2305.16291v1 [cs.AI]. DOI: 10.48550/arXiv.2305.16291. arXiv: 2305.16291 [cs.AI]. URL: https://arxiv.org/abs/2305.16291.

[15] Shunyu Yao et al. "REACT: SYNERGIZING REASONING AND ACTING IN LANGUAGE MODELS". In: *arXiv preprint arXiv:2210.03629* (2022). URL: https://arxiv.org/pdf/2210.03629.pdf.

10 Appendix

Figure 22: Citrine AGI Agent Running Live on Twitch

Figure 23: Adventure Mode for ChatWorld

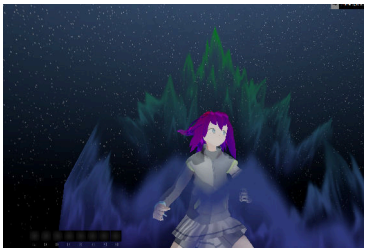

Figure 24: A game engine with interactive avatars

Figure 25: Imaginative story dialogue

Figure 26: A procedurally generated game world

Figure 27: In-engine lighting effects 1

Figure 28: In-engine lighting effects 2

Figure 29: In-world day-night cycle

Figure 30: In-world custom avatars

10.1 Image gallery

Figure 31: Animatediff can generate compelling custom animations

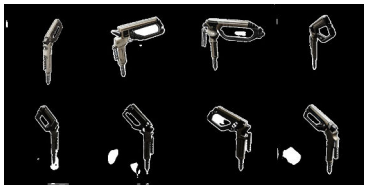

Figure 32: 3D items generated with stable diffusion and Zero-123

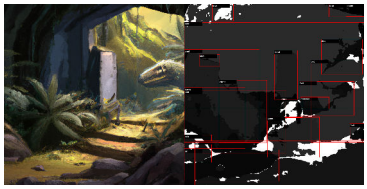

Figure 33: Image segmentation allows for visual understanding of worlds

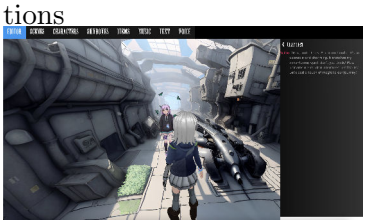

Figure 34: Blockade labs depth-based skyboxes can generate infinite linked worlds

Figure 35: Embodied VRM characters can exist in Blockade Labs worlds

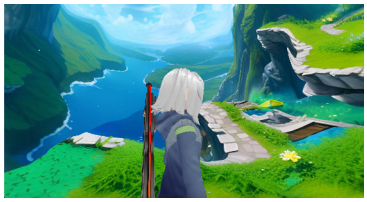

Figure 36: The scale of Blockade Labs worlds can be vast

Figure 37: Avatars can wear objects in the world

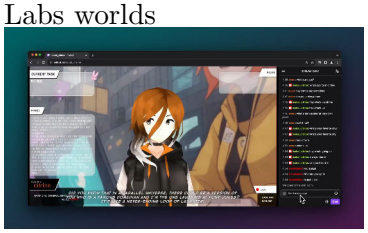

Figure 38: "Citrine" artificial general intelligence agent on Twitch 1

Figure 39: "Citrine" artificial general intelligence agent on Twitch 2

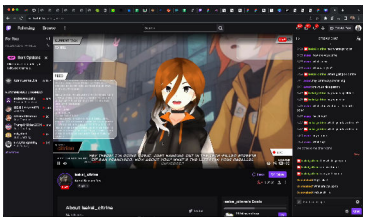

Figure 40: "Citrine" artificial general intelligence agent on Twitch 3

Figure 41: Items can be generated with Midjourney 1

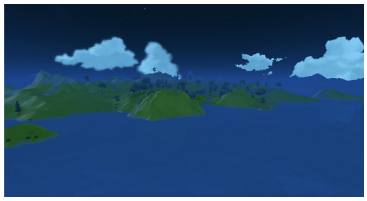

Figure 42: Procedurally generated world base

Figure 43: The engine supports character facial expressions

Figure 44: Image segmentation can be used to perceive the world

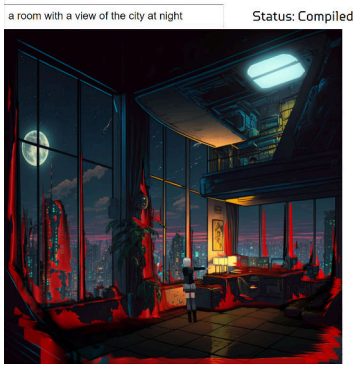

Figure 45: Images depth can be used to place characters into artwork

Figure 46: Stable diffusion character generation 1

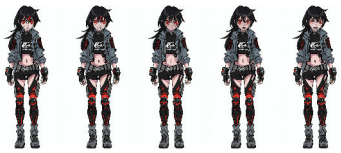

Figure 47: Stable diffusion character generation 2

Figure 48: Stable diffusion character generation 3

Figure 49: Character base image generation

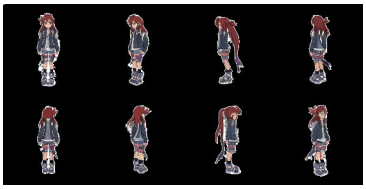

Figure 50: Character 360 spritesheet generation

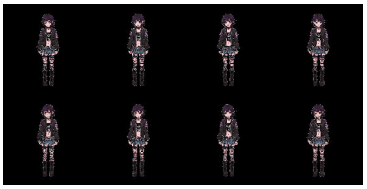

Figure 51: Character expressions spritesheet generation

Figure 52: Guided stable diffusion 1

Figure 53: Guided stable diffusion 2

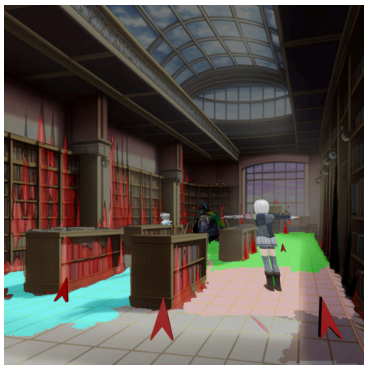

Figure 54: World generation can support scene occlusion

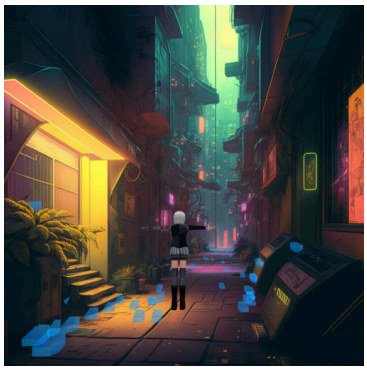

Figure 55: Image depth extrustion can support a large range of depths

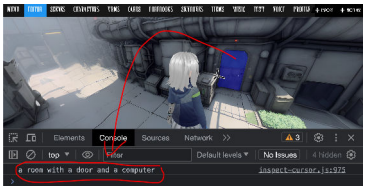

Figure 56: SAM + BlIP models can be used for in-world understanding

Figure 57: Stable diffusion character generation 1

Figure 58: Stable diffusion character generation 2

Figure 59: Stable diffusion character generation 3

Figure 60: Stable diffusion character generation 4

Figure 61: Stable diffusion character generation 5

Figure 62: Particle effects

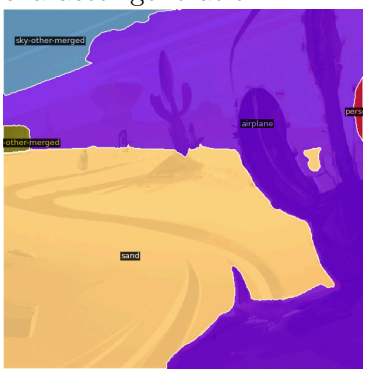

Figure 63: Image segmentation over art

Figure 64: 2D images can be processed into 3D worlds

Figure 65: 2D -¿ 3D comic panel processing pipeline 1

Figure 66: 2D -¿ 3D comic panel processing pipeline 2

Figure 67: 2D -¿ 3D comic panel processing pipeline 3

Figure 68: 2D -¿ 3D comic panel processing pipeline 4

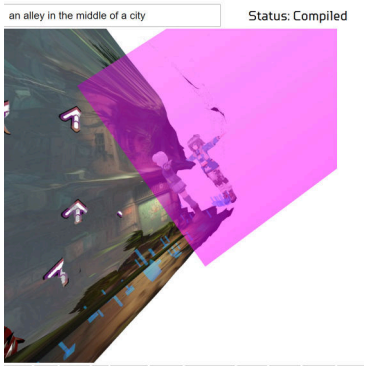

Figure 69: 2D -¿ 3D image pipeline editing

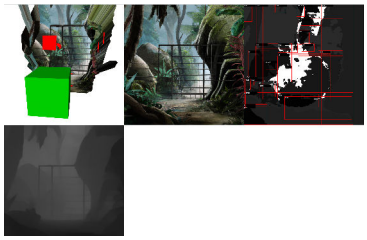

Figure 70: 2D -¿ 3D image processing pipeline

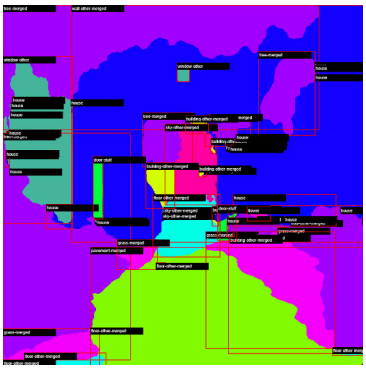

Figure 71: 2D scene segmentation

Figure 72: Procedurally generated story scenes 1

Figure 73: Procedurally generated story scenes 2

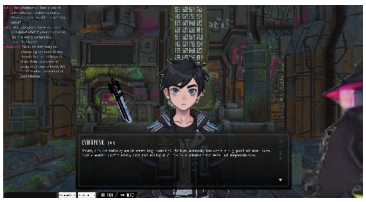

Figure 74: Procedurally generated story scenes 3

Figure 75: Realtime speech input in engine

Figure 76: Characters scene

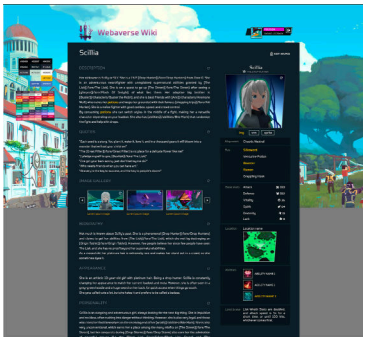

Figure 77: Procedurally generated wiki 1

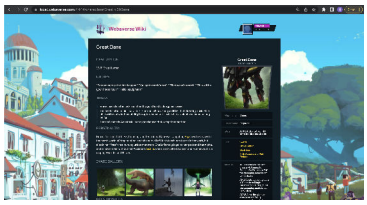

Figure 78: Procedurally generated wiki 2

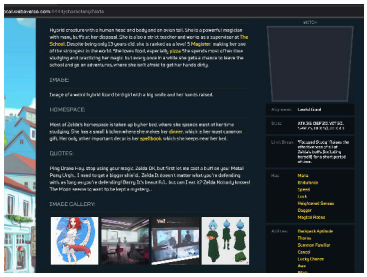

Figure 79: Procedurally generated wiki 3