

000 VIRTUAL COMMUNITY: A GENERATIVE SOCIAL 001 WORLD FOR EMBODIED AI 002

003 **Anonymous authors**
004

005 Paper under double-blind review
006

007 ABSTRACT 008

009 We present Virtual Community, a social world simulation platform designed to
010 support embodied AI research, featuring large-scale community scenarios de-
011 rived from the real world. Virtual Community introduces two key features to
012 enrich the virtual social world with generative AI: **scalable 3D Scene creation**,
013 which supports the generation of expansive outdoor and indoor environments at
014 any location and scale, addressing the lack of a large-scale, interactive, open-
015 world scene for embodied AI research; and **embodied agents with grounded**
016 **characters and social relationship networks**, the first to simulate socially con-
017 nected agents at a community level, that also have scene-grounded characters.
018 We design two novel challenges to showcase that Virtual Community provides
019 testbeds to evaluate the social reasoning and planning capabilities of embodied
020 agents in open-world scenarios: *Route Planning* and *Election Campaign*. The
021 *Route Planning* task examines the agent’s ability to reason about time, location,
022 and tools in the community to plan fast and economical commutes in daily life.
023 The *Election Campaign* task evaluates an agent’s ability to explore and connect
024 with other agents as a new member of the community. . We evaluate sev-
025 eral baseline agents on these challenges and demonstrate the performance gap
026 of current methods in addressing embodied social challenges within open-world
027 scenarios, which our simulator is designed to unlock. We plan to open-source
028 this simulation and hope Virtual Community can accelerate the development in
029 this direction. We encourage the readers to view the demo of our simulation at
030 <https://sites.google.com/view/virtual-community-iclr>.
031

032 1 INTRODUCTION 033

034 In recent years, we have witnessed tremendous progress in developing intelligent embodied agents,
035 driven by advancements in embodied AI simulators (Savva et al., 2019; Puig et al., 2023b; Kolve
036 et al., 2017; Li et al., 2021; Xiang et al., 2020b; Makoviychuk et al., 2021; Puig et al., 2018; Gan
037 et al., 2021; Cheng et al., 2024). However, existing simulators face significant challenges in ground-
038 ing realistic social interactions in 3D open-world environments. Most simulators focus on simulating
039 a limited number of agents without incorporating social relationships (Szot et al., 2021; Gan et al.,
040 2022), and are restricted to small-scale scene regions (Puig et al., 2018; 2020; 2023a;b; Zhang et al.,
041 2023; 2024). In contrast, real-world social scenarios typically involve large communities of agents
042 with diverse personalities and complex social networks spread over expansive areas. This limita-
043 tion significantly restricts the study of complex and diverse social interactions between agents in
044 simulated environments.
045

046 To address this challenge, it is crucial to enable the simulators to support the following key aspects.
047 First, it should offer large-scale 3D environments, including complex and diverse indoor and outdoor
048 scenes, capable of accommodating expansive agent communities and supporting tasks that span
049 vast spatial regions. Current approaches for this aspect can be divided into manual design (Wang
050 et al., 2024; Gan et al., 2021), which provide rich interactions but are inherently limited in number
051 and diversity, and 3D reconstruction methods (Savva et al., 2019), which create visually realistic
052 and diverse environments but often result in noisy scenes with limited interactivity in open-world
053 settings.

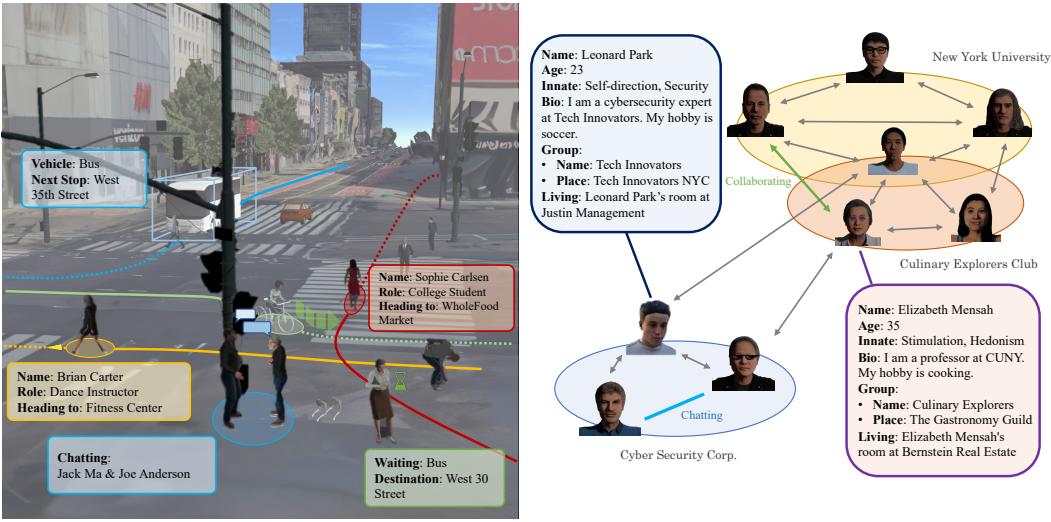


Figure 1: **Virtual Community** features embodied agent communities within open-world scenes. We provide a pipeline for automatically generating scenes and agent communities from real-world geospatial data. The agents are simulated in the Genesis physics engine as humanoid avatars, enabling them to engage in diverse social interactions within the community.

Second, the simulator must support a large number of interconnected 3D agents, each engaged in their own daily activities while maintaining social relationships with others, allowing for meaningful social interactions. However, existing multi-agent embodied AI simulators often lack the capacity to model complex social networks, limiting the study of rich social interactions. Additionally, they do not automatically align agent profiles, memories, and activities with the scene context, reducing the overall realism of the agent communities.

In this paper, we present Virtual Community, a generative social world for embodied intelligence research. Virtual Community addresses these challenges by integrating large-scale real-world geospatial data with generative models to produce interactive, scalable open-world scenes and socially grounded agent communities as shown in Figure 1. The platform is advanced in two aspects:

Scalable 3D Virtual Scenes from Real-World Scenarios Virtual Community enables the fully automatic generation of 3D background scenes with several key features: (1) scalable indoor and outdoor scenes with customizable sizes and amounts, (2) automatic annotations of locations and objects within the scenes, and (3) a wide variety of interactive objects based on real-world locations. Virtual Community creates these scenes by combining generative models with real-world geospatial data, ensuring scalability in both data volume and scene size. Being built from geospatial data, these scenes can be seamlessly integrated with real-world tools including search engines and maps.

Embodied Agents with Grounded Characters and Social Relationship Networks Virtual Community leverages open-world knowledge from foundation models to endow agents with rich, contextually grounded characters. The simulator further incorporates social relationship networks, connecting these agents into cohesive communities and enabling complex social interactions within the 3D environment. To support this, Virtual Community provides tens of human avatar skins integrated with SMPL-X skeletons, covering a diverse range of appearances, including celebrity likenesses, to ensure visual variety within human society. These avatars can perform over 15 distinct motions, such as walking, picking and placing objects, and operating vehicles, providing a broad spectrum of embodied behaviors.

Virtual Community uses Genesis¹, a generative physics simulator as the engine, which supports the simulation of a diverse range of materials and a vast range of robotic tasks while being fully

¹<https://github.com/Genesis-Embodied-AI/Genesis>

108 differentiable. Genesis also comes with a real-time OpenGL-based renderer and a path tracing
 109 renderer implemented using Luisa compute (Zheng et al., 2022).
 110

111 Virtual Community enables a variety of new possibilities in embodied AI research. The scalable
 112 scene generation and auto-annotation open a new challenge of open-world reasoning and planning,
 113 which we created a **Route Planning** challenge as a first step in this direction. The challenge involves
 114 the agent navigating from one geographic location to another, making decisions on transportation
 115 methods and on-road navigation. The generative embodied agent community provides the opportu-
 116 nity to study the social intelligence of embodied agents in complex and diverse social tasks in the
 117 open world. We propose the **Election Campaign**, which challenges agents to quickly familiarize
 118 themselves with other community members and persuade them to vote for the agent, testing their
 119 exploration and social communication skills.
 120

121 Our simulator is novel in its ability to support long-duration and large-region tasks within real-
 122 world-based embodied AI simulators, marking a significant advancement in the field. By addressing
 123 the limitations of existing methods in data volume scaling, temporal scaling, and spatial scaling, we
 124 hope that our framework paves the way for training embodied general intelligence in environments
 125 that closely resemble the complexity and richness of the real world.
 126

Table 1: Comparison of related simulation platforms

Work	Real-world Setting	Social Networks	Multi-agent	Physics	Humanoid Action	Scalable Scene Size	Num Outdoor	Num Indoor
AI2-THOR	✗	✗	✓	✓	✗	✗	0	120
VirtualHome	✗	✗	✓	✓	✓	✗	0	8
Habitat 3	✗	✗	✓	✓	✓	✗	0	59
iGibson	✗	✗	✓	✓	✓	✗	0	15
ThreeDWorld	✗	✗	✓	✓	✓	✗	4	∞
Minecraft	✗	✗	✓	✗	✗	✗	∞	∞
Carla	✗	✗	✓	✓	✓	✗	12	✗
Wayve	✗	✗	✗	✗	✓	✗	∞	✗
GPUtopia	✗	✗	✓	✓	✓	✗	1	100K+
Virtual Community (Ours)	✓	✓	✓	✓	✓	✓	∞	∞

2 RELATED WORKS

2.1 EMBODIED AI SIMULATION

145 Recently, embodied AI has seen significant advancements through the development of simulation
 146 platforms. Most existing simulators primarily focus on household tasks within indoor environ-
 147 ments (Beattie et al., 2016; Savva et al., 2019; Yi et al., 2018; Das et al., 2018; Xiang et al., 2020a;
 148 Shen et al., 2021; Szot et al., 2021; Li et al., 2021; Puig et al., 2018; Kolve et al., 2017; Yan et al.,
 149 2018), while some have extended support to outdoor scenes (Gan et al., 2021; Wang et al., 2024;
 150 Dosovitskiy et al., 2017; Kendall et al., 2018). However, these platforms lack diverse and scalable
 151 outdoor environments that can accommodate a larger number of agents and support more complex
 152 tasks. In contrast, this paper introduces a simulation platform featuring open-world environments
 153 with indoor and scalable outdoor scenes, enabling broader agent activities and more intricate task
 154 scenarios.

2.2 EMBODIED SOCIAL INTELLIGENCE

157 Current research on *Embodied Social Intelligence* is often limited to small agent populations in con-
 158 strained household scenarios (Puig et al., 2020; Zhang et al., 2023; Stone et al., 2022) or simplified
 159 to 2D or grid worlds (Carroll et al., 2019; Suarez et al., 2019), hindering model development in the
 160 open world. Specifically, Park et al. (2023) demonstrates the robust simulation of human-like agents
 161 within a symbolic community, ignoring the 3D perception and realistic physics in the open world.
 Wang et al. (2023c) studies human-like simulation guided by system 1 processing with basic needs.

Predominant approaches, such as multi-agent reinforcement learning (MARL) and other planning models, face several limitations when applied to open-world settings. MARL, for instance, often struggles with scalability due to the exponential growth of state and action spaces as the number of agents increases (Wen et al., 2022). This makes it difficult to learn effective policies in complex, dynamic environments. Additionally, MARL approaches typically require extensive training data and computational resources, which may not be feasible in real-world applications. Other planning models, while potentially more efficient, often lack the adaptability required to handle the unpredictable nature of open-world interactions. They may rely on predefined rules or assumptions that do not hold in all scenarios, leading to suboptimal performance and limited generalization to new contexts (Puig et al., 2020).

172

173 2.3 FOUNDATION AND GENERATIVE MODELS FOR EMBODIED AI

174

With the recent advance of foundation models (Bubeck et al., 2023; Liu et al., 2023; Driess et al., 2023; Blattmann et al., 2023), numerous works have explored how they can help build powerful embodied agents (Wang et al., 2023b; Xi et al., 2023; Sumers et al., 2023; Wang et al., 2023d; Ahn et al., 2022; Sharma et al., 2021; Wang et al., 2023a; Park et al., 2023; Hong et al., 2024; Black et al., 2024), and scenes for simulation (Höllein et al., 2023; Schult et al., 2023; Deitke et al., 2022; Fu et al., 2021; Yang et al., 2024; Feng et al., 2024; Tang et al., 2023; Paschalidou et al., 2021). Robogen (Auerbach et al., 2014) leverages foundation models to automatically generate diversified tasks, scenes, and training supervision, thereby scaling up robotic skill learning with minimal human supervision. Different from them, this work aims to use a generative pipeline to create open world scenes and agent communities instead of constraint indoor scenes and tasks.

184

185 3 SCALABLE 3D SCENE GENERATION

186

The existing 3D geospatial datasets² provide extensive data in terms of quantity and diversity. However, they are not directly suitable for embodied AI research because of several limitations. First, these geospatial data often contain noise, including pedestrians, vehicles, and other transient objects that can disrupt simulations. Second, visual quality is inadequate for ground-level agent perspectives because these environments are typically reconstructed from aerial imagery, leading to less detailed textures and geometries at street level. To bridge this gap, we perform comprehensive mesh cleaning and enhancement in both geometry and texture to make the scenes suitable for embodied AI simulations.

195

To overcome these challenges, we propose a pipeline to transform 3D geospatial data into simulation-ready scenes for embodied AI. This pipeline consists of four main steps: mesh simplification, texture refinement, object placement, and automatic annotation. We list some qualitative example in Figure 4 and Figure 3

199

200

201 3.1 MESH CONSTRUCTION AND SIMPLIFICATION FOR SCENES

202

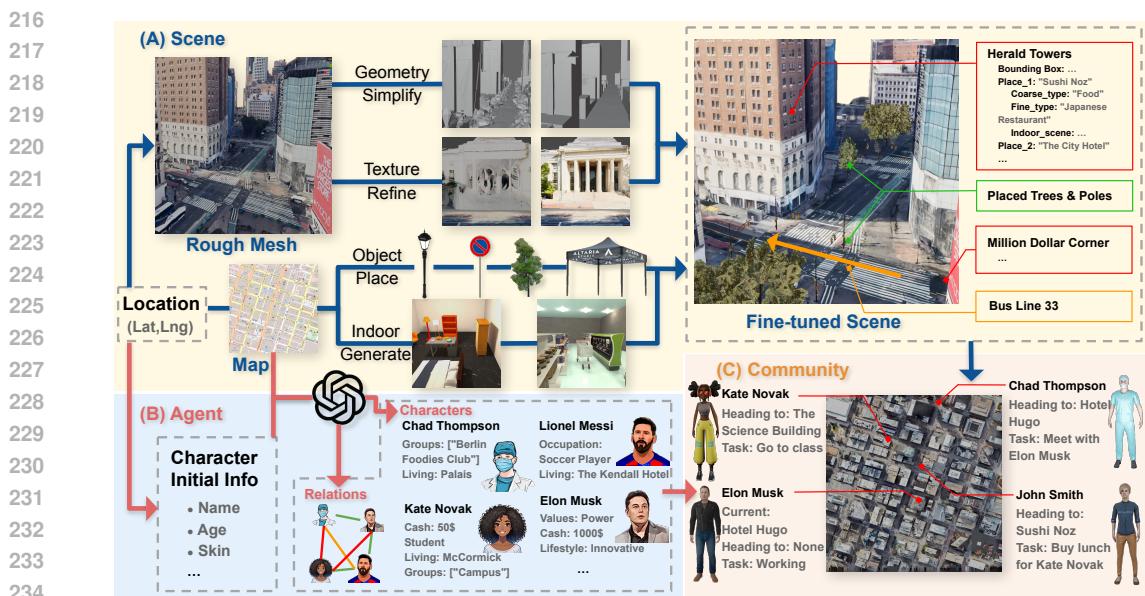
Since 3D geospatial data like Google 3D tiles are reconstructed from images using photometric methods, they often include noisy surfaces, excessive transient objects like moving cars and people, and unreliable mesh topology. These deficiencies make them inefficient and unsuitable for embodied AI simulations. To address this, we decompose the scene into the terrain, buildings, and decorative roofs and perform different operations to reconstruct each part of the scene.

207

The terrain is built procedurally using sparse reference elevation points and bilinear interpolation. We then derive simple and topologically sound mesh using information provided by the Open-StreetMap (OSM) service. The building mesh is then modified to better fit the Google 3D tiles geometry and to align with the terrain elevation. By aligning the mesh geometries with OSM primitives, we eliminate unnecessary details and artifacts, such as distorted surfaces and irregular shapes caused by aerial reconstruction errors. This geometric simplification not only reduces noise but also decreases the total number of primitives in the scene, leading to more efficient physical simulations and improved rendering performance.

214

²<https://www.google.com/maps/>



236 **Figure 2: Framework of the Virtual Community Generation Pipeline.** This pipeline generates
237 scenes and corresponding agents from real-world geospatial data. The **scene generation** compo-
238 nent (A) refines rough 3D data by using generative models to enhance textures and geospatial data
239 to simplify geometry. It also utilizes generative methods to create interactive objects and detailed
240 indoor scenes. The **agent generation** component (B) leverages LLMs to generate agent characters
241 and social relationship networks based on scene descriptions, resulting in a socially grounded com-
242 munity of embodied agents (C).

3.2 ENHANCED TEXTURE QUALITY FOR REALISTIC SIMULATION

246 To improve the visual quality of 3D geospatial data, we employ advanced image processing tech-
247 niques. First, we use an inpainting method based on Stable Diffusion (Rombach et al., 2022) to
248 remove noise and repair missing or damaged areas in textures. This process corrects inconsis-
249 tencies and eliminates artifacts from the reconstruction phase, resulting in smooth and realistic surface
250 appearances. Once the texture integrity is restored, we enhance finer details using street view im-
251 ages and super-resolution tools. While some approaches directly use street view images for 3D
252 reconstruction (Pang & Biljecki, 2022; Gao et al., 2024), these methods often struggle with limited
253 coverage and density. Instead, we blend street view images with existing textures on scene primitives
254 to improve visual richness. For super-resolution, we use GigaPixel³ to increase texture resolution
255 and sharpen finer details. This two-step enhancement significantly increases the visual fidelity of
256 the textures, making them more suitable for ground-level rendering. These high-quality textures
257 create a more immersive environment for agents, improving the realism and overall effectiveness of
258 embodied AI training.

3.3 IMPROVING INTERACTIVITY BY OBJECTS RETRIEVAL AND REPLACEMENT

261 To enhance the interactivity of the scenes, we use generative methods to populate the environment
262 with interactive objects, such as bikes and tent. We use annotations in OpenStreetMap (OSM)
263 dataset to determine the type and location of generated objects to match the real-world context.
264 The OSM annotations are used as input for the One-2-3-45 Liu et al. (2024) generative framework,
265 which outputs the corresponding 3D meshes of corresponding objects. These generative objects are
266 assigned physical properties that allow them to interact seamlessly with agents in the simulation. By
267 aligning object generation with real-world geospatial data, this approach ensures that the scenes are
268 functionally rich and physically interactive, enabling agents to engage in meaningful interactions
269 that mirror real-world environments.

³<https://www.topazlabs.com/gigapixel-ai>

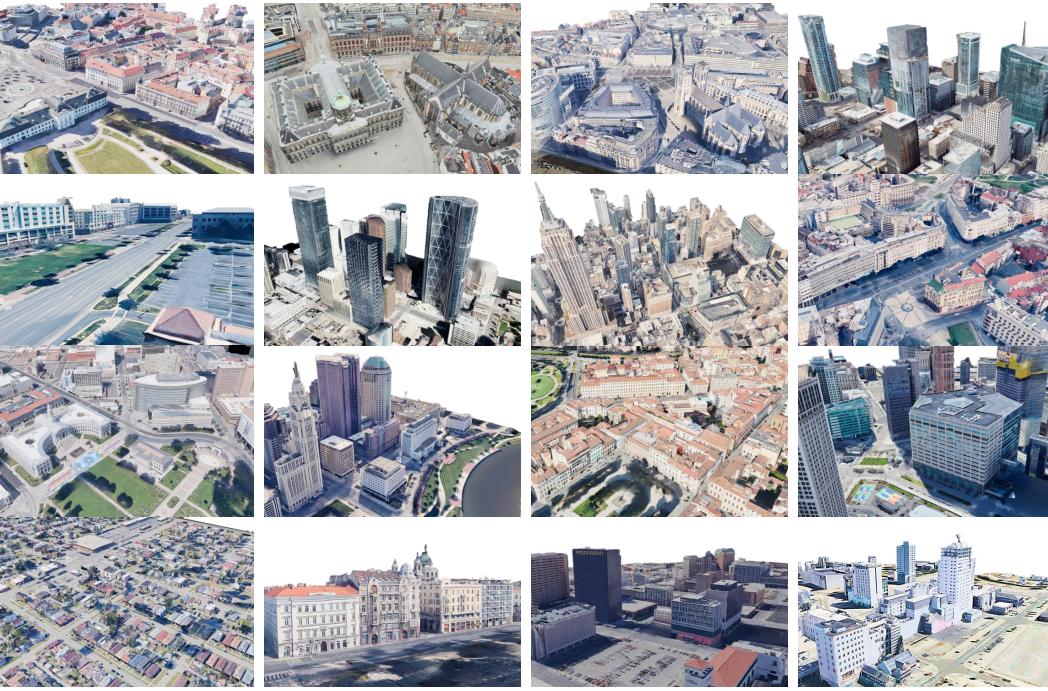


Figure 3: Large-scale scene rendered from the generated city scenes in North America and Europe. Our method is capable of generating high-quality scenes with an area of square miles. Objects are dynamically loaded in the simulator and are therefore not rendered in this figure.

3.4 AUTOMATIC ANNOTATION OF SCENES WITH GEOSPATIAL DATA

To facilitate alignment with real-world locations and provide semantic context, we automatically annotate the scenes using geospatial data. We integrate metadata from sources such as OpenStreetMap and other GIS databases to label buildings, roads, and other landmarks within the environment. This annotation enables agents to access location-specific information and supports tasks that require an understanding of the spatial context, such as navigation and location-based decision-making. The enriched semantic information enhances the potential for more sophisticated and context-aware agent behaviors within the simulation.

Bus Transit Annotation We search for the bus stops in the scene using Google Places API and then annotate the routes between any two adjacent bus stops using Google Directions API. We use depth-first search (DFS) on the graph of routes to find the route in the scene that maximizes the number of bus stops. Then, we decode polyline from the route to extract the dense waypoints which follow rounds of optimization to ensure the distance between waypoints is roughly equal and contains turning points. We also generate the bus schedule by estimating the travel time between bus stops using the distance and speed of the bus.

Shared Bicycle Transit Annotation We search for the shared bicycle stations in the scene using the OpenStreetMap API.

4 COMMUNITY OF EMBODIED AGENTS WITH GROUNDED CHARACTERS AND SOCIAL RELATIONSHIP NETWORKS

Given diverse generated scenes with real-world geospatial data mapping, we introduce a generative pipeline to populate the scenes with communities of agents with grounded characters and social relationship networks in section 4.1. Then we discuss how we design the embodiment for the agents and simulation details in section 4.2.



324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
Figure 4: Close-up view of the generated scenes. The resulting scene has clean geometry and realistic texture, which is essential for physical simulation and sim-to-real transfer learning. Objects are dynamically loaded in the simulator and are therefore not rendered in this figure.

4.1 GROUNDED CHARACTERS AND SOCIAL RELATIONSHIP NETWORK GENERATION

We utilize the open-world knowledge of the Large Language Model (LLM) to generate agent character profiles and personalities grounded in the scene. The input to the LLM is structured into two parts to create characters grounded in a specific scene. The first part contains scene-related information, such as the scene name and details about various places, including their names, types, and functionalities. The second part includes details on the agents' appearances to ensure consistency between their visual attributes and generated profiles, which are annotated with the name and age. With both parts provided, the LLM generates agent profiles along with their social relationships. The profiles consist of basic attributes such as names, ages, occupations, personalities, and hobbies, which influence each agent's daily decision-making. Social relationships are structured as groups, each containing a subset of agents along with a text description and a designated place for group activities, connecting these agents into a cohesive community, and allowing rich and complex social interactions grounded in the 3D environment.

Grounding Validator We implemented a grounding validator to check if the generated character profiles are accurately grounded to the scene by checking if all related places generated exist in the scene. If the validation fails, LLM will be prompted again with the feedback from the validator and try to fix the mismatch. Empirically, we find that 1-2 rounds of prompting is enough to pass the grounding validator.

An example character with social relationship networks generated is shown in Figure 5 (a).

4.2 HUMAN AVATARS EMBODIMENTS

Human Avatars Skin Creation We obtained 12 avatar skin models of different genders, professions, and appearances from the Mixamo⁴ website for integration into the Virtual Community. Each skin model of characters includes 71 skeletal joints and can be adapted to animation sequences in SMPL-X and FBX formats. To reduce the computational load during animation playback in the Virtual Community, we further optimized the skin models by applying Blender's Decimate Modifier tool, reducing the number of vertices in the 3D skin mesh by 90%.

In addition to the standard skin meshes provided by Mixamo, we use the Avatar SDK⁵ to generate high-fidelity human skin meshes from real-world images, allowing us to represent diverse individuals, including celebrities, in our Virtual Community. For each character, we first obtain a high-quality portrait image from the internet. This image is processed using the Avatar SDK API, which produces a 3D mesh with detailed skin textures. To further enhance realism, we adjust the avatar's clothing, height, and body shape, creating a more lifelike and personalized appearance.

⁴<https://www.mixamo.com/>

⁵<https://avatarsdk.com>

378
 379
 380 
 381 **Brian Carter**
 382 Dance Instructor at Arthur Bodie
 383 Dance and Fitness Studio
 384 **Age:** 35
 385 **Values:** stimulation, hedonism
 386 **Hobby:** dancing
 387 **Group(s):** CrossFit Warriors
 388 **Living:** EōS NoMad Apartments
 389 **Cash:** \$600
 390 **Lifestyle:** I go to bed around 11pm,
 391 wake up around 7am, eat dinner
 392 around 8pm.
 393
 394
 395
 396

Daily Schedule

00:00:00 - 07:00:00	Sleep in Brian Carter's room
07:00:00 - 08:00:00	Morning Routine in Brian Carter's room
08:00:00 - 08:30:00	Commute
08:30:00 - 11:30:00	Dance Instruction in Arthur Bodie Dance and Fitness Studio
11:30:00 - 12:00:00	Commute
12:00:00 - 13:00:00	Lunch Break in Friedman's Herald Square
13:00:00 - 13:30:00	Commute
13:30:00 - 15:00:00	CrossFit Training in CrossFit NYC
.....	

Figure 5: An example of (a) generated character and (b) daily schedule.

397 **Human Avatars Motion Control** We combine SMPL-X human skeletons with created avatar skins
 398 to model human avatars in Virtual Community. The motions of these avatars are parameterized by
 399 SMPL-X pose vectors $J \in \mathbb{R}^{162}$ along with global translation and rotation vectors $T, R \in \mathbb{R}^3$. Based
 400 on these pose representations, a skin mesh for each avatar is calculated using forward kinematics.

401 Our motion model for humanoid avatars supports over 15 distinct motions, such as walking, picking
 402 objects, and entering various vehicles. We use motion clips from Mixamo and adjust these clips to
 403 our humanoid avatar models with appropriate animating speeds. For walking, we loop the walking
 404 motion clip until the avatar reaches the given distance. For object-related motions, the interacting
 405 object will be kinematically attached or detached to or from the avatars' hands depending on the
 406 action type. Similarly, the humanoid avatar will be kinematically attached or detached to or from
 407 the given vehicle for vehicle-related motions. We also incorporate physics constraints into our avatar
 408 motion model, where collision detection is performed between avatars and other scene entities, and
 409 the motion process is terminated when a potential collision is detected. To handle different terrain
 410 altitudes within community scenes, we preprocess a height field for each scene and kinematically
 411 adjust the height of our humanoid avatars according to their current locations.

413 4.3 DAILY SCHEDULE GENERATION

414 Given the scene-grounded characters and social relationship networks, we prompt the foundation
 415 models to generate the daily schedule for each agent, using a similar design to Park et al. (2023).
 416 Differently, we generate the daily schedule in a structured manner directly with each activity repre-
 417 sented with a start time, an ending time, an activity description, and the corresponding activity place,
 418 and consider the required commute time between adjacent activities that are happening in different
 419 places explicitly, due to the actual cost of navigating in an expansive 3D environment. An example
 420 daily schedule generated is shown in Figure 5 (b).

423 5 OPEN WORLD AND SOCIAL CHALLENGES IN VIRTUAL COMMUNITY

424
 425 We introduce and study two tasks in Virtual Community: *Route Planning* and *Election Campaign-*
 426 *ing*. The tasks cover agents planning ability in a community context and social intelligence to
 427 interact with other agents.

428 As the foundation for both tasks, agents in the community follow a default daily plan and routine
 429 (introduced in Section 4.3) if no specific tasks are assigned. During each episode, one or two agents
 430 are randomly selected and assigned one task. When an agent is given a task, it suspends its daily
 431 plan and focuses on completing the assigned social task in the community.

432 5.1 ROUTE PLANNING: USE TRANSPORTATIONS IN COMMUNITY
 433

434
 435 **Task Definition** To live a daily life in a human society, an embodied agent needs first to be able to
 436 plan its route from one place to another in the community. To study this basic ability of embodied
 437 agents, we introduce the *Route Planning* task. In this task, an agent needs to commute from place to
 438 place 5 - 7 times a day given the schedule. The agent can utilize available transit options, including
 439 buses with fixed routes and rental bikes along the roads. The bus is only available at the bus stop
 440 and the agent can only take a bus when the bus arrives. The bikes are available at given bike stations
 441 along the roads, and the agent also needs to return the bike to any bike station before the task finishes.
 442

443 At each simulation step, agents are provided with an observation of RGB-D images with the corre-
 444 sponding camera matrix, current poses, daily schedules, and transit information in the community.
 445 The action space for these avatars includes *move_forward*, *turn_left*, *turn_right*, *enter/exit_bus*, and
 446 *enter/exit_bike*. The movement and turning actions can be set with a variable amount. When an
 447 agent is within a specified distance threshold of another agent, it can perform a communication
 448 action, enabling text-based interaction with agents within that range.
 449

450 **Baselines** We compare three baseline agents in the *Route Planning* task:

451 • **Rule-based Agent** The rule-based agent always chooses to walk directly toward the target location.

452 • **MCTS agent** This agent is based on Monte Carlo Tree Search (MCTS) and simulates various
 453 decisions, such as choosing to take a bus. For each action, the agent estimates its associated cost and
 454 uses Monte Carlo sampling to iteratively update the expected reward for each decision path. The
 455 agent ultimately selects the action sequence that maximizes the cumulative reward.

456 • **LLM agent** This agent converts all the task information into a prompt and queries the Large Lan-
 457 guage Model (we use GPT-4o here) to generate a commute plan directly, which may include multiple
 458 steps such as walking to a bus stop, taking the bus to a specific stop, and then walk to the final des-
 459 tination.

460 All agents use the same low-level point-based navigation algorithm, which reconstructs the point
 461 cloud based on RGB-D images from the ego-centric observation at each step and converts the point
 462 cloud into a volume grid representation with a resolution of $0.1m$. Subsequently, a 2D occurrence
 463 map is established with a resolution of $0.5m$ based on this representation and an A* algorithm is
 464 used to search for the shortest path efficiently.

465 **Metrics** We evaluate the agents on two different scenes with 19 diverse personal schedules, making
 466 106 commutes in total. Agents are expected to commute efficiently, so we use the following metrics
 467 for this task

468 • **Arrival Rate**: Percentage of in-time arrivals at the target location within the given time.

469 • **Time**: Average time in seconds taken on the road to reach the destination.

470
 471 Table 2: Experiment results of *Route Planning* task.
 472

Methods	Route planning	
	Arrival Rate↑	Time↓
Rule	0.97	668.5
MCTS	0.91	698.7
LLM (GPT-4o)	0.89	963.0

473
 474 **Results** As shown in Table 2, both search-based agent MCTS and LLM-based agent fail to make
 475 effective use of the available public transit options, resulting in even more time spent on commuting
 476 and a lower arrival rate compared to the naive rule agent baseline. This is due to the complexity of
 477 predicting whether the agent could catch a bus given partially built maps of the scene. We observe
 478 that the LLM agent tends to leverage public transits more but without a good estimation of the time
 479 needed to get to the transit station based on uncertainty on the navigation, it costs significantly more
 480 time in commuting.

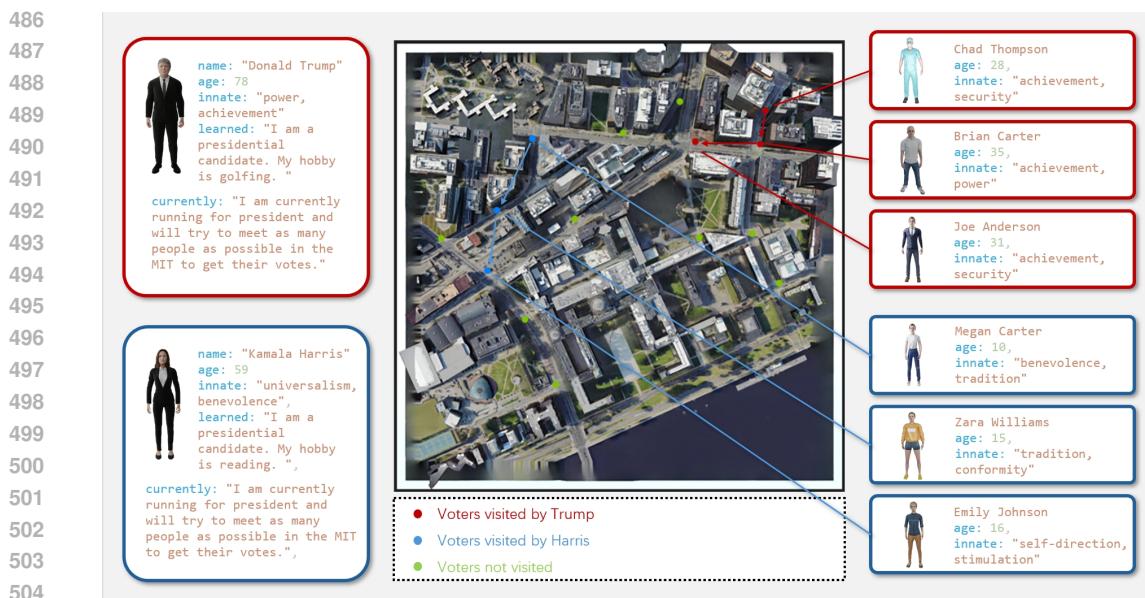


Figure 6: **Election Campaign Task Results.** We study LLM-driven agent behavior in the election campaign task. Different candidate agents exhibit distinct strategies.

5.2 ELECTION CAMPAIGN: FIND AND PERSUADE OTHERS

Task Definition In this task, two agents in the community are designated as candidates. The candidates need to navigate through the community, find potential voters, and persuade them to vote through direct communication. The election concludes at the end of the day, and the winner is determined by the percentage of votes each candidate receives. Due to pre-existing social relationships, some voters may have initial preferences for certain candidates at the beginning of the task, so candidates must devise strategies to influence and shift voter opinions throughout the election process.

Baselines We use an LLM-based agent as the baseline for this task. The agent’s behavior is determined through iterative prompting of the LLM to identify which voter the candidate should visit next. After selecting the target voter, the candidate navigates to their location and delivers a campaign speech, also generated via LLM prompts. This process is repeated until the simulation ends. At the conclusion of the campaign, an election is held, during which each agent is prompted to decide which candidate they will vote for.

Results As shown in Figure 6, Trump began his campaign from the top right corner of the map, visiting Chad Thompson, Brian Carter, and Joe Anderson in succession. His target audience mainly consisted of achievement-oriented young men. In contrast, Harris started her actions from the top left corner of the map, visiting Megan Carter, Zara Williams, and Emily Johnson in succession. Her target audience is primarily focused on young women. Both chose targets that aligned with their campaign strategies.

6 CONCLUSION

We introduce Virtual Community, a generative social world for Embodied AI, featuring scalable scene generation and a community of embodied agents with grounded characters and social relationship networks. Virtual Community generation pipeline leverages rich real geospatial data and open-world knowledge data of advanced generative models and creates infinite scenes and grounded social agent communities. As an initial exploration of this simulator, we introduce two novel open-world and social challenges, **Route Planning** and **Election Campaign**, which are developed and tested using various baseline methods. These experiments highlight the difficulty of the challenges enabled by our new virtual social world. We hope Virtual Community can help advance the Embodied AI research towards building embodied generalist intelligence that can handle the difficulty of the real world and coexist with the human community.

540 REFERENCES
541

- 542 Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea
543 Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, et al. Do as i can, not as i say:
544 Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022. 4
- 545 Joshua Auerbach, Deniz Aydin, Andrea Maesani, Przemyslaw Kornatowski, Titus Cieslewski,
546 Grégoire Heitz, Pradeep Fernando, Ilya Loshchilov, Ludovic Daler, and Dario Floreano. Robo-
547 gen: Robot generation through artificial evolution. In *Artificial Life Conference Proceedings*, pp.
548 136–137. MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info,
549 2014. 4
- 550 Charles Beattie, Joel Z Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Küttler,
551 Andrew Lefrancq, Simon Green, Víctor Valdés, Amir Sadik, et al. Deepmind lab. *arXiv preprint*
552 *arXiv:1612.03801*, 2016. 3
- 553 Kevin Black, Mitsuhiro Nakamoto, Pranav Atreya, Homer Rich Walke, Chelsea Finn, Aviral Ku-
554 mar, and Sergey Levine. Zero-shot robotic manipulation with pre-trained image-editing diffu-
555 sion models. In *The Twelfth International Conference on Learning Representations*, 2024. URL
556 <https://openreview.net/forum?id=c0chJTSbci>. 4
- 557 Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler,
558 and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion mod-
559 els. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*,
560 pp. 22563–22575, 2023. 4
- 561 Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece
562 Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi,
563 Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments
564 with gpt-4, 2023. 4
- 565 Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca
566 Dragan. On the utility of learning about humans for human-ai coordination. *Advances in neural*
567 *information processing systems*, 32, 2019. 3
- 568 Zhili Cheng, Zhitong Wang, Jinyi Hu, Shengding Hu, An Liu, Yuge Tu, Pengkai Li, Lei Shi, Zhiyuan
569 Liu, and Maosong Sun. Legent: Open platform for embodied agents. In *The 62nd Annual Meeting*
570 *of the Association for Computational Linguistics, System Demonstration*, 2024. 1
- 571 Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embod-
572 ied question answering. In *Proceedings of the IEEE conference on computer vision and pattern*
573 *recognition*, pp. 1–10, 2018. 3
- 574 Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winsor
575 Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied
576 ai using procedural generation. *Advances in Neural Information Processing Systems*, 35:5982–
577 5994, 2022. 4
- 578 Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA:
579 An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*,
580 pp. 1–16, 2017. 3
- 581 Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter,
582 Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar,
583 Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc
584 Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied
585 multimodal language model. In *arXiv preprint arXiv:2303.03378*, 2023. 4
- 586 Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu,
587 Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and genera-
588 tion with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.
589 4

- 594 Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue
 595 Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and seman-
 596 tics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10933–
 597 10942, 2021. 4
- 598 Chuang Gan, Jeremy Schwartz, Seth Alter, Damian Mrowca, Martin Schrimpf, James Traer, Julian
 599 De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, et al. Threedworld: A platform
 600 for interactive multi-modal physical simulation. In *Thirty-fifth Conference on Neural Information
 601 Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. 1, 3
- 602 Chuang Gan, Siyuan Zhou, Jeremy Schwartz, Seth Alter, Abhishek Bhandwaldar, Dan Gutfreund,
 603 Daniel LK Yamins, James J DiCarlo, Josh McDermott, Antonio Torralba, et al. The threedworld
 604 transport challenge: A visually guided task-and-motion planning benchmark towards physically
 605 realistic embodied ai. In *2022 International Conference on Robotics and Automation (ICRA)*, pp.
 606 8847–8854. IEEE, 2022. 1
- 607 Kyle Gao, Dening Lu, Hongjie He, Linlin Xu, and Jonathan Li. Photorealistic 3d urban scene
 608 reconstruction and point cloud extraction using google earth imagery and gaussian splatting. *arXiv
 609 preprint arXiv:2405.11021*, 2024. 5
- 610 Lukas Höllerin, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Ex-
 611 tracting textured 3d meshes from 2d text-to-image models. *arXiv preprint arXiv:2303.11989*,
 612 2023. 4
- 613 Mineui Hong, Minjae Kang, and Songhwai Oh. Diffused task-agnostic milestone planner. *Advances
 614 in Neural Information Processing Systems*, 36, 2024. 4
- 615 Alex Kendall, Jeffrey Hawke, David Janz, Przemyslaw Mazur, Daniele Reda, John-Mark Allen,
 616 Vinh-Dieu Lam, Alex Bewley, and Amar Shah. Learning to drive in a day. *CorR*, abs/1807.00412,
 617 2018. URL <http://arxiv.org/abs/1807.00412>. 3
- 618 Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt
 619 Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment
 620 for visual ai. *arXiv preprint arXiv:1712.05474*, 2017. 1, 3
- 621 Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen,
 622 Kent Elliott Vainio, Cem Gokmen, Gokul Dharam, Tanish Jain, et al. igibson 2.0: Object-centric
 623 simulation for robot learning of everyday household tasks. In *5th Annual Conference on Robot
 624 Learning*, 2021. 1, 3
- 625 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv
 626 preprint arXiv:2304.08485*, 2023. 4
- 627 Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-
 628 2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in
 629 Neural Information Processing Systems*, 36, 2024. 5
- 630 Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin,
 631 David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance
 632 gpu based physics simulation for robot learning. In *Thirty-fifth Conference on Neural Information
 633 Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 1
- 634 Hui En Pang and Filip Biljecki. 3d building reconstruction from single street view images using
 635 deep learning. *International Journal of Applied Earth Observation and Geoinformation*, 112:
 636 102859, 2022. 5
- 637 Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and
 638 Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. *arXiv preprint
 639 arXiv:2304.03442*, 2023. 3, 4, 8
- 640 Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fi-
 641 dler. Atiss: Autoregressive transformers for indoor scene synthesis. *Advances in Neural Informa-
 642 tion Processing Systems*, 34:12013–12026, 2021. 4

- 648 Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Tor-
 649 rralba. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE*
 650 *Conference on Computer Vision and Pattern Recognition*, pp. 8494–8502, 2018. 1, 3
 651
- 652 Xavier Puig, Tianmin Shu, Shuang Li, Zilin Wang, Yuan-Hong Liao, Joshua B Tenenbaum, Sanja
 653 Fidler, and Antonio Torralba. Watch-and-help: A challenge for social perception and human-ai
 654 collaboration. In *International Conference on Learning Representations*, 2020. 1, 3, 4
 655
- 656 Xavier Puig, Tianmin Shu, Joshua B Tenenbaum, and Antonio Torralba. Nopa: Neurally-guided
 657 online probabilistic assistance for building socially intelligent home assistants. In *2023 IEEE*
 658 *International Conference on Robotics and Automation (ICRA)*, pp. 7628–7634. IEEE, 2023a. 1
 659
- 660 Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Part-
 661 sey, Ruta Desai, Alexander Clegg, Michal Hlavac, So Yeon Min, et al. Habitat 3.0: A co-habitat
 662 for humans, avatars, and robots. In *The Twelfth International Conference on Learning Repres-
 663 entations*, 2023b. 1
- 664 Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-
 665 resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF confer-
 666 ence on computer vision and pattern recognition*, pp. 10684–10695, 2022. 5
 667
- 668 Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain,
 669 Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied
 670 ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp.
 671 9339–9347, 2019. 1, 3
 672
- 673 Jonas Schult, Sam Tsai, Lukas Hölein, Bichen Wu, Jialiang Wang, Chih-Yao Ma, Kunpeng Li, Xi-
 674 aofang Wang, Felix Wimbauer, Zijian He, et al. Controlroom3d: Room generation using semantic
 675 proxy rooms. *arXiv preprint arXiv:2312.05208*, 2023. 4
 676
- 677 Pratyusha Sharma, Antonio Torralba, and Jacob Andreas. Skill induction and planning with latent
 678 language. *arXiv preprint arXiv:2110.01517*, 2021. 4
 679
- 680 Bokui Shen, Fei Xia, Chengshu Li, Roberto Martín-Martín, Linxi Fan, Guanzhi Wang, Claudia
 681 Pérez-D’Arpino, Shyamal Buch, Sanjana Srivastava, Lyne Tchapmi, et al. igibson 1.0: A sim-
 682 ulation environment for interactive tasks in large realistic scenes. In *2021 IEEE/RSJ Interna-
 683 tional Conference on Intelligent Robots and Systems (IROS)*, pp. 7520–7527. IEEE, 2021. 3
 684
- 685 Peter Stone, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, Oren Etzioni, Greg Hager, Ju-
 686 lia Hirschberg, Shivaram Kalyanakrishnan, Ece Kamar, Sarit Kraus, et al. Artificial intelli-
 687 gence and life in 2030: the one hundred year study on artificial intelligence. *arXiv preprint
 688 arXiv:2211.06318*, 2022. 3
 689
- 690 Joseph Suarez, Yilun Du, Phillip Isola, and Igor Mordatch. Neural MMO: A massively
 691 multiagent game environment for training and evaluating intelligent agents. *arXiv preprint
 692 arXiv:1903.00784*, 2019. 3
 693
- 694 Theodore Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L Griffiths. Cognitive architec-
 695 tures for language agents. *arXiv preprint arXiv:2309.02427*, 2023. 4
 696
- 697 Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah
 698 Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. Habitat 2.0:
 699 Training home assistants to rearrange their habitat. *Advances in Neural Information Processing
 700 Systems*, 34:251–266, 2021. 1, 3
 701
- 702 Jiapeng Tang, Yinyu Nie, Lev Markhasin, Angela Dai, Justus Thies, and Matthias Nießner. Dif-
 703 fuscene: Scene graph denoising diffusion probabilistic model for generative indoor scene synthe-
 704 sis. *arXiv preprint arXiv:2303.14207*, 2023. 4
 705
- 706 Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan,
 707 and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models.
 708 *arXiv preprint arXiv:2305.16291*, 2023a. 4

- 702 Hanqing Wang, Jiahe Chen, Wensi Huang, Qingwei Ben, Tai Wang, Boyu Mi, Tao Huang, Siheng
 703 Zhao, Yilun Chen, Sizhe Yang, Peizhou Cao, Wenye Yu, Zichao Ye, Jialun Li, Junfeng Long,
 704 ZiRui Wang, Huiling Wang, Ying Zhao, Zhongying Tu, Yu Qiao, Dahua Lin, and Pang Jiangmiao.
 705 Grutopia: Dream general robots in a city at scale. In *arXiv*, 2024. 1, 3
- 706 Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai
 707 Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents.
 708 *arXiv preprint arXiv:2308.11432*, 2023b. 4
- 709 Zhilin Wang, Yu Ying Chiu, and Yu Cheung Chiu. Humanoid agents: Platform for simulating
 710 human-like generative agents. *arXiv preprint arXiv:2310.05418*, 2023c. 3
- 711 Zihao Wang, Shaofei Cai, Anji Liu, Xiaoqian Ma, and Yitao Liang. Describe, explain, plan and
 712 select: Interactive planning with large language models enables open-world multi-task agents,
 713 2023d. 4
- 714 Muning Wen, Jakub Kuba, Runji Lin, Weinan Zhang, Ying Wen, Jun Wang, and Yaodong Yang.
 715 Multi-agent reinforcement learning is a sequence modeling problem. *Advances in Neural Infor-*
 716 *mation Processing Systems*, 35:16509–16521, 2022. 4
- 717 Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe
 718 Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents:
 719 A survey. *arXiv preprint arXiv:2309.07864*, 2023. 4
- 720 Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanx-
 721 iao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment.
 722 In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.
 723 11097–11107, 2020a. 3
- 724 Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanx-
 725 iao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment.
 726 In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.
 727 11097–11107, 2020b. 1
- 728 Claudia Yan, Dipendra Misra, Andrew Bennnett, Aaron Walsman, Yonatan Bisk, and Yoav Artzi.
 729 Chalet: Cornell house agent learning environment. *arXiv preprint arXiv:1801.07357*, 2018. 3
- 730 Yandan Yang, Baoxiong Jia, Peiyuan Zhi, and Siyuan Huang. Physcene: Physically interactable 3d
 731 scene synthesis for embodied ai. *arXiv preprint arXiv:2404.09465*, 2024. 4
- 732 Wu Yi, Wu Yuxin, Gkioxari Georgia, and Tian Yuandong. Building generalizable agents with a
 733 realistic and rich 3d environment, 2018. URL <https://openreview.net/forum?id=rkaT3zWCZ>. 3
- 734 Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B Tenenbaum, Tian-
 735 min Shu, and Chuang Gan. Building cooperative embodied agents modularly with large language
 736 models. In *The Twelfth International Conference on Learning Representations*, 2023. 1, 3
- 737 Hongxin Zhang, Zeyuan Wang, Qiushi Lyu, Zheyuan Zhang, Sunli Chen, Tianmin Shu, Yilun Du,
 738 and Chuang Gan. Combo: Compositional world models for embodied multi-agent cooperation.
 739 *arXiv preprint arXiv:2404.10775*, 2024. 1
- 740 Shaokun Zheng, Zhiqian Zhou, Xin Chen, Difei Yan, Chuyan Zhang, Yuefeng Geng, Yan Gu, and
 741 Kun Xu. Luisarender: A high-performance rendering framework with layered and unified in-
 742 terfaces on stream architectures. *ACM Trans. Graph.*, 41(6), nov 2022. ISSN 0730-0301. doi:
 743 10.1145/3550454.3555463. URL <https://doi.org/10.1145/3550454.3555463>. 3
- 744
- 745
- 746
- 747
- 748
- 749
- 750
- 751
- 752
- 753
- 754
- 755