Dynamic Scene Generation for Embodied Navigation Benchmark

Chenxu Wang, Dong Wang, Xinghang Li, Dunzheng Wang, Huaping Liu*

***Abstract*—Although embodied agents have been widely studied in multifarious tasks within abundant benchmarks, studies in dynamic scenarios have not been sufficiently supported by large-scale dynamic scenes. Many existing works aim at probabilistic environments where the daily object may be moved due to human activities, however, the scale of the dataset is usually limited due to the cost of human annotation or manual configuration. Toward the scalable generation of such dynamic scenes, we introduce a framework that simulates human activities and corresponding object dynamics with Large Language Models (LLMs) and apply the simulated human residents to embodied scenes. A user study that compares our generated scene dynamics with other approaches validates that our framework successfully produces believable and diversified data, which have a quality comparable to human annotations. We further conduct object goal navigation experiments under various problem settings with representative baselines on dynamic scenes. The results verify the potential of generated scenes to serve as navigation benchmarks while suggesting that dynamic scenes introduce new challenges and problems to embodied navigation. Our work contributes as an infrastructure that may facilitate future studies on embodied AI in dynamic environments. A visualization and online demonstration of our framework dynamic scene generation is available at https://huggingface.co/spaces/JW0003/DynamicSceneGeneration**

I. Introduction

Beyond the static environments that have widely been adopted in the study of Embodied AI, dynamic scenes where the daily object may be moved according to human behaviors have also attracted a lot of attention. The probabilistic nature of dynamic scenes introduces more challenges, including constructing object-based memory system [1], proactive robot assistance [2, 3], object-goal navigation with a probabilistic object configuration [4, 5], and navigation within environments where exist dynamic humanoid obstacles [6].

However, most existing studies on dynamic scenes conduct evaluation either in human-annotated environments with only a few scenes or in simplified 2D environments, limiting the generalizability of the results. On the other hand, although embodied simulators have achieved fruitful success [7, 8, 9, 10], and large-scale scenes can be generated in various ways [11, 12, 13], those approaches remain generating static scenes, where no human activity influence and object dynamic are considered. To support future embodied AI research in dynamic scenes and fill in the gap of lacking large-scale dynamic scene datasets, we introduce a framework for generating dynamic scenes by simulating human behaviors and configuring corresponding object relocations with time in static scenes to obtain dynamic scenes. Object positions in dynamic scenes will evolve with time according to human activities, as illustrated in Figure 1.

For automatic and scalable simulating human indoor activities, we seek the aid of Large Language Models (LLMs), which have been widely utilized in human simulation [14, 15] and have a verified power of content generation. Our framework first hierarchically generates human activity schedules based on customizing persona, then generates a set of probabilities of object relocations caused by the activities according to the designated human persona. The simulated human avatars are then assigned to specific static scenes, configuring the object relocations and ending up with dynamic scenes. Trading off between scalability and simulation quality, we keep the human simulation agnostic to specific scene layouts, resulting in simulated human behaviors are applicable to all scenes in the same domain, for example, the 10,000 scenes in the ProcTHOR dataset [11].

To evaluate our framework and the quality of generated dynamic scenes, we create 50 characters and simulate 5-day activities for each character in the context of the ProcTHOR dataset. A user study confirms that the quality of our generated data is comparable to human annotations, and even better in comprehensiveness and diversity. To validate the potential of dynamic scenes as a benchmark for embodied navigation tasks, we demonstrate object-goal navigation tasks in dynamic scenes in various problem settings with representative baselines. The results also suggest that dynamic scenes introduce new challenges and room for future object goal navigation tasks.

The main contributions of this paper are summarized as follows:

- We introduce a hierarchical framework for simulating human indoor activities with LLMs and scalable generation of dynamic scenes by configuring the object relocations in static scenes.
- We build an extensible dynamic scene dataset based on the ProcTHOR dataset and conduct a user study to verify the quality of the simulated human behaviors in dynamic scenes.
- We demonstrate several settings of object-goal navigation tasks with corresponding baselines. Our data generation framework and public-available data may facilitate future studies on various tasks in the dynamic scenario.

The authors are with Department of Computer Science and Technology, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, China. This work was jointly supported by the National Natural Science Fund for Distinguished Young Scholars under Grant 62025304 and National Natural Science Foundation Project under Grant 62273054. * denotes the corresponding author: hpliu@tsinghua.edu.cn

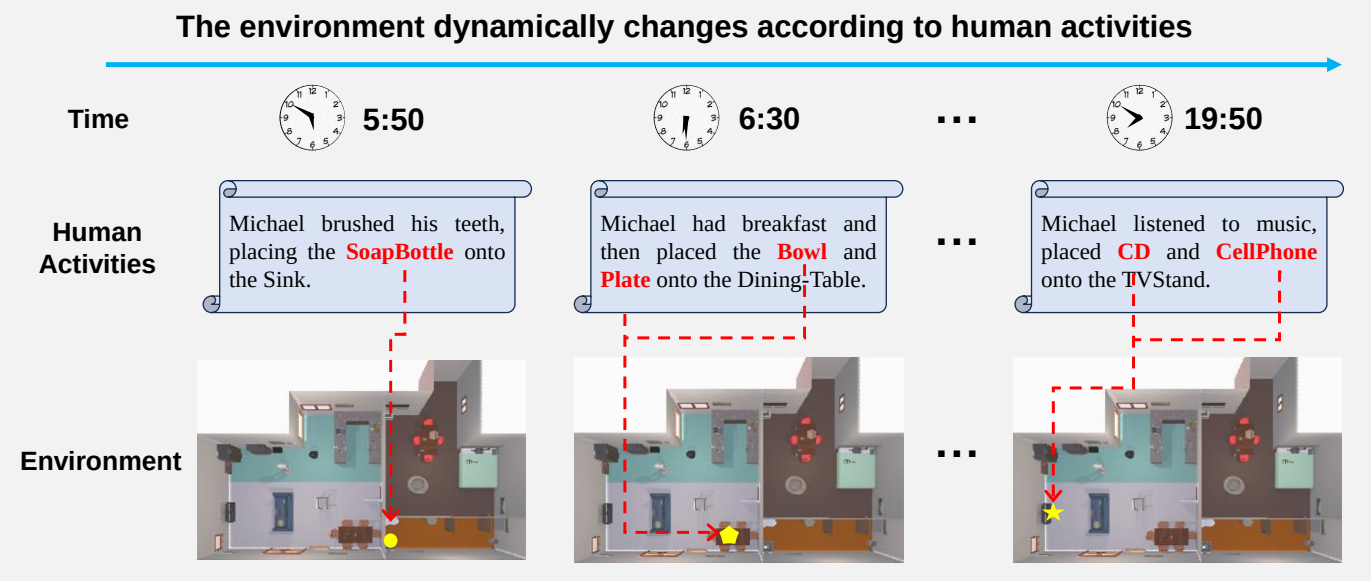


Fig. 1: An illustration of the idea of the dynamic scene, where the daily object may be moved over time according to human activities. The relations between human activities and the objects are highlighted with the red dashed lines and the yellow marks denote the designation place of object relocations.

II. Related Work

A. Embodied Scene Datasets

Along with the widespread attention of researchers on embodied AI, embodied simulators have been extensively developed [7, 16, 17, 18, 19, 20]. Driven by the need for large-scale scene datasets, ProcTHOR procedurally generates over 10,000 houses that are applicable for 3D navigation tasks. Previous studies validate the feasibility of generating photorealistic scenes [12], indoor scenes [21, 22], data for visual-language navigation tasks [13], and even physically interactable scenes [23]. With the support of abundant data, studies on embodied tasks such as object goal navigation have achieved substantial progress [24, 25, 26, 27, 28].

Despite the success of static scenarios, embodied AI studies in dynamic environments also require the support of datasets, however, existing benchmarks are rather preliminary. [1] uses environments with manually configured object relocation patterns for learning and evaluating object-based memory systems. [5] studies map-based object goal navigation on two simulated and one real office kitchen environments under probabilistic goal configurations. Beyond manually designing the probabilities, [4] proposed iGridson, a 2D environment that samples the object placement according to a scene graph where the object relations can be extracted from large-scale 3D datasets. A concurrent work, DOZE [6], contains 10 scenes with moving humanoid obstacles to evaluate the agents' collision avoidance abilities in low-level navigation. We regard the HOMER series [2, 3] as the most related work, which collects fine-grained human annotations of various daily activities with corresponding object interactions in the VirtualHome environment [29]. However, despite the high quality of manual annotation, the collected data is hard to generalize to new scenes with different layouts, thus the scale of the dataset is limited by the economic cost. To generate dynamic scenes on a large scale, we utilize the power of LLMs to simulate human activities and configure the object dynamics on existing static scenes, making our framework as scalable as static scene generation.

B. Human Activity Simulation

Modeling and synthesizing human behaviors have also been recognized as a remarkable task in robotic research. Based on the collected human activity data, [2] also introduces a framework for the procedural synthesizing of new human avatars. However, the synthetic human data are still limited in scale and are difficult to generalize to scenes with different layouts. Idrees et al. [30] introduces a procedural framework for simulating human daily activity based on manually created schedule templates, however, the requirement of human workload limits its applicability in generating large-scale datasets.

Fortunately, recent studies have substantiated that LLMs are capable of simulating humans in various domains and synthesizing data for downstream tasks [31, 32, 33]. For instance, [14] and [34] study the capability of LLMs in simulating human samples and replicating human subject studies. Another thread of research validates the capabilities of LLM-based agents to effectively mimic human-like social behavior [35, 36] and produce meaningful results in several application domains such as diplomacy [37] and job fair [38]. Besides, LLMs can also be used as a human user simulator to train downstream models [39]. Beyond fidelity, benefiting from its capability of role-playing [40, 41], utilizing LLMs in simulating humans can also offer diversity for data. Our work is also inspired by the generative agents [15], which presents a believable human simulation in a sandbox simulator,

including both hierarchical schedule planning and action-level interaction with the environment, directly substantiating the feasibility of simulating both human behaviors and the subsequent environment dynamics. Nevertheless, the aforementioned works still focus on the text domain, in which context our work contributes a believable, economical, and extensible framework of dynamic scene generation with LLM-based human simulation to facilitate future embodied navigation studies.

III. Problem Formulation

In this work, we aim to generate dynamic scenes where the object placements vary with time according to the influence of human activities. Specifically in the indoor domain, a dynamic scene $\mathcal{S}$ can be defined as a tuple:

$$\mathcal{S} = (\mathcal{H}, \mathcal{L}, \mathcal{R}, \mathcal{O}, \mathcal{P}, t_s, t_e),$$

where $\mathcal{H}$ stands for the static house layout, $\mathcal{L}$ is the set of rooms, $\mathcal{R}$ is the set of immovable receptacles, $\mathcal{O}$ is the set of movable objects, $\mathcal{P} : \mathcal{O} \times [t_s, t_e] \mapsto \mathcal{R}$ is the configuration of object positions over time, t_s is the start time of the dynamic scene, and t_e is the end time. We further denote the dynamic scene at time t as $\mathcal{S}_t = (\mathcal{H}, \mathcal{L}, \mathcal{R}, \mathcal{O}, \mathcal{P}_t)$, where $\mathcal{P}_t(\cdot) = \mathcal{P}_t(\cdot, t)$. Embodied task can be set at any $\mathcal{S}_t$ within the constraint $t_s \leq t \leq t_e$.

Recall that the primary cause of scene dynamic is the influence of human activities, thus the evolution of the scene is defined by events, which can be represented as $e = (t_s^{(e)}, t_e^{(e)}, a^{(e)}, \mathcal{P}^{(e)})$, where $t_s^{(e)}$ and $t_e^{(e)}$ stand for the starting and ending time instant of the event, $a^{(e)}$ is a natural language description for the event, and $P^{(e)}$ is the object relocations caused by the event, which is a set of object-receptacle pairs. Each dynamic scene is associated with a set of events, denoted as $\mathcal{E}$. Given a static scene with $\mathcal{H}, \mathcal{L}, \mathcal{R}, \mathcal{O}$ and a given time interval (t_s, t_e), we aim to generate a reasonable event set $\mathcal{E}$ and subsequently generate the object configuration $\mathcal{P}$.

IV. Generating Dynamic Scenes by Simulating Human Activities

In this section, we introduce our framework for human resident simulation and subsequent dynamic scene generation. As illustrated in Figure 1, our framework takes pre-defined character information and domain-specific scene prior information as input, followed by three modules: generation of the human activity schedules; establishment of the activity database that models the environment evolution caused by each activity; and a scene configuration module that applies the object relocations to specific scenes to get the event set $\mathcal{E}$. Both the generation of the schedule and the activity database are powered by LLMs [1].

The external input data can be in pure textual form, where the characters are defined by 8 fields, dividing into basic information and characteristic information as shown in Figure 2 (a). Character personalities can be customized by manual designation or automatically generated as well. Learning from the success of SOTOPIA [35], we include 50 characters that are initialized by GPT-4 with manual filtering and adjustment to avoid duplication and enrich the diversity. The scene prior information offers a profile of the target simulation domain, including the possible rooms, receptacles, objects, and possible relations.

It is noteworthy that our framework is designed in a modularized way where the activity schedules and the activity database are independent of specific scenes, enabling character to reuse in different scenes. The amount of distinct characters determines the diversity and comprehensiveness of our dataset, while the data can be grounded in infinite scenes.

A. Human Activity Schedule

Aiming to provide believable and personalized activity schedules, we draw on the idea of hierarchical planning from generative agents [15] and implement a top-down framework for human schedule generation, which first generates a general plan and then decompose the indoor parts into detailed activities, as illustrated in Figure 2 (b).

The general plan is intended to reflect the characteristics of the persona, planning the day at the hour level with personalized sketchy description, such as *doing research* or *preparing for a party*. At this stage, only the character information is provided.

Despite the colorful general plan, we ultimately aim at building indoor dynamic scenes. Hence the general plan is subsequently filtered by substituting the outside part with *leave home* to force the LLM to focus on the indoor activities. The filtered plan is then decomposed into detailed activity schedules, where the activities are more concrete and at the minute level, such as *getting dressed* and *reading the newspaper*. To guide the LLM, we additionally provide an activity list as examples and a reference for granularity. Such an activity list extends continuously in the generation process by merging with new activities, initializing by a predefined list of daily activities acquired from the HOMER dataset [2].

B. Activity Database

After acquiring the human activity schedules, the natural subsequent step is to analyze how the activities impose effects on the environment. To this end, we build an activity database that hierarchically models the probabilities of object relocation caused by each activity as illustrated in Figure 2 (c), considering the characteristics of the persona.

We assume an activity will be performed at only one location. Given an activity a, we denote the possibility of it being performed at locations l as $p(l|a) \in [0, 1]$, with the constraint $\sum_{l \in \mathcal{L}} p(l|a) = 1$ for any activity a. The probabilities are under comprehensive consideration of the nature of activities, characteristics of the persona, and the available options of locations with possible receptacles that reflect the functionality of the locations.

Once activities are paired with specific locations, we inspect the probabilities of objects being used or involved in each

[1]We use gpt-4-1106-preview in the dynamic scene generation process.

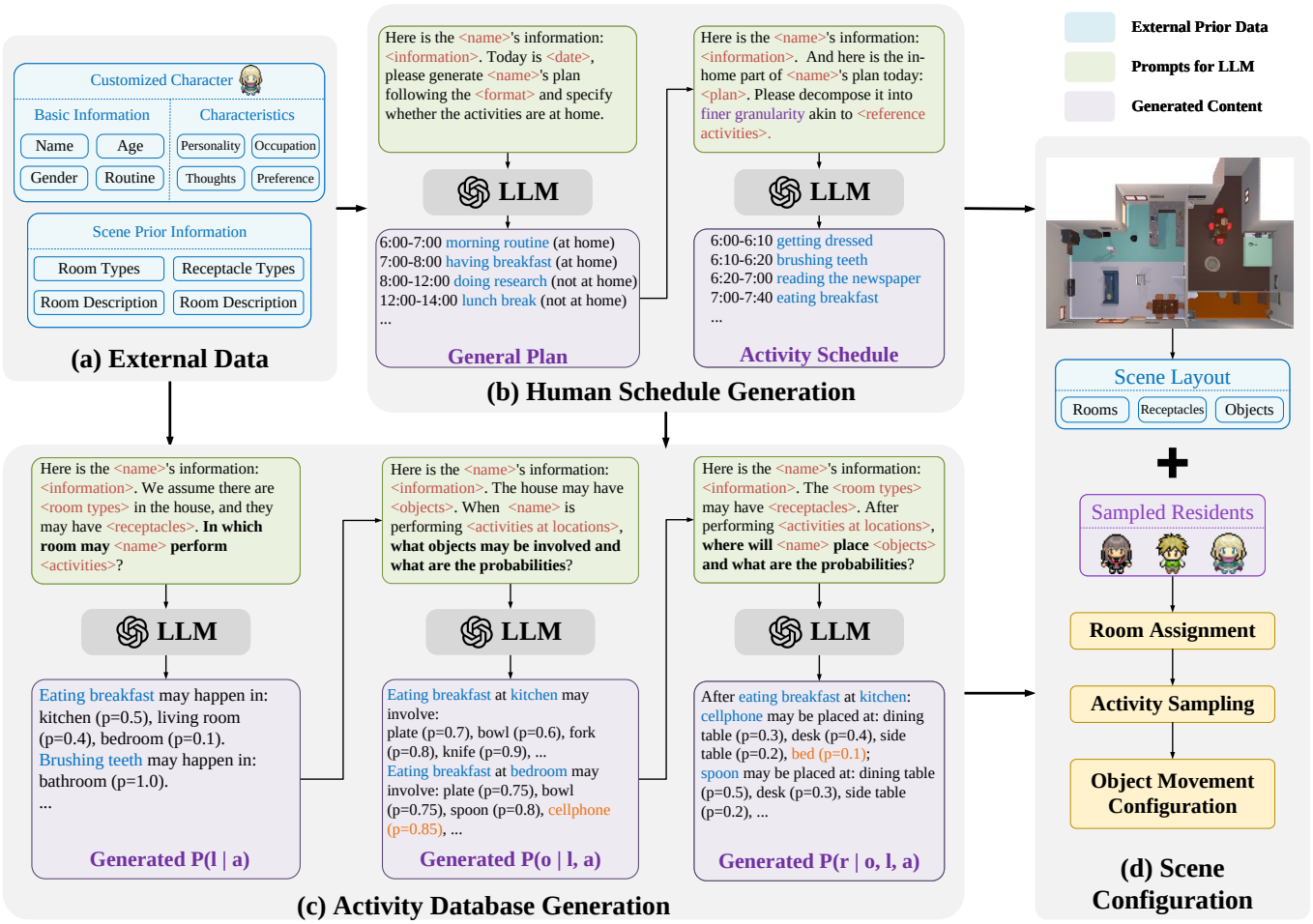


Fig. 2: An overview of the dynamic scene generation framework. Given the information of characters and the scene prior, our framework first generates the activity schedules and the corresponding activity database to simulate a resident avatar. Static scenes can be extended to dynamic ones by sampling residents and configuring the object relocations caused by their activities.

event, and which objects are considered to be moved due to the effect of the activities. For example, tableware may be used for *having lunch at kitchen* and subsequently placed on the *dining table*. Then, after *tidying up the dining table at kitchen*, they may be placed back to the *countertop*. In this regard, we use the probability $p(o|a,l) \in [0,1]$ to characterize the possibility of the object o being involved in activity a at location l.

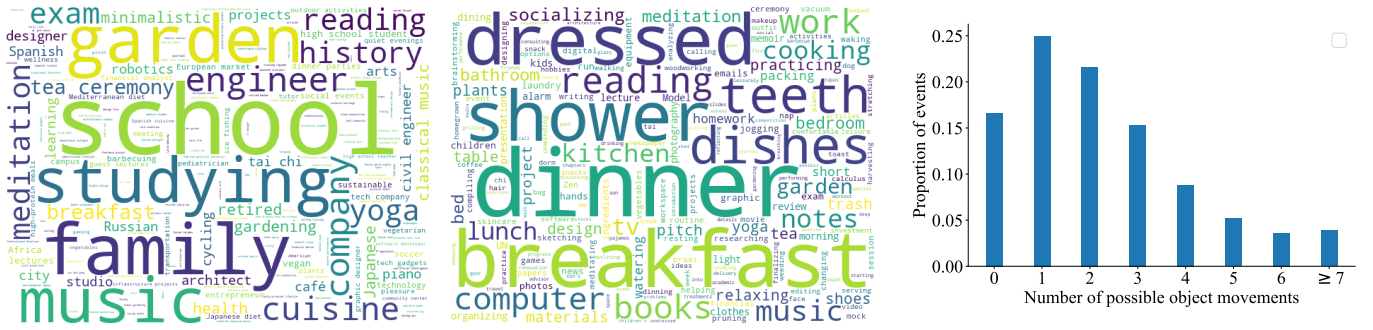
For some object o, if the probability $p(o|a,l) > 0$, i.e., it may be involved in the activity, then we use $p(r|o,a,l)$ to denote the probability of the human moving an object with type o onto a receptacle $r \in \mathcal{R}$. We assume that a used object will be placed onto exactly one receptacle. This results in $\sum_{r \in \mathcal{U}} p(r|o,a,l) = 1$.

C. Scene Configuration

As shown in Figure 2 (d), the final step of our framework is to ground the human activities into specific scenes to instantiate the scene evolution by forming the event set $\mathcal{E}$ and calculate the object position configuration $\mathcal{P}$.

The configuration process is tightly coupled with the scene layout. In this paper, we build dynamic scenes based on the initial scenes provided by the ProcTHOR dataset [11], including four types of rooms: bedroom, bathroom, living room, and kitchen, while the number of rooms may vary in different scenes. We start by assuming each person lives in a unique bedroom and first sample residents according to the number of bedrooms in the scene. With a given time interval, the event set $\mathcal{E}$ can be formed by incorporating all indoor activities of all sampled residents.

For each event in $\mathcal{E}$, we sample the effect following the hierarchical probabilistic model described earlier: first sampling the location of the activities, followed by the objects involved and the target receptacles for the used objects. In consideration of realism, residents are assigned specific bedrooms and bathrooms, *i.e.*, they will always choose their assigned room whenever they want to go to a bedroom or a bathroom. Since the activity database is generated with consideration of all possible receptacles, the relocation probabilities will be first normalized by the existence of receptacles in the given scene.



(a) Word cloud of the persona characteristics. (b) Word cloud of the generated activities. (c) Proportion of events potentially involving different numbers of objects.

Fig. 3: Visualization and statistics of our simulated human activities.

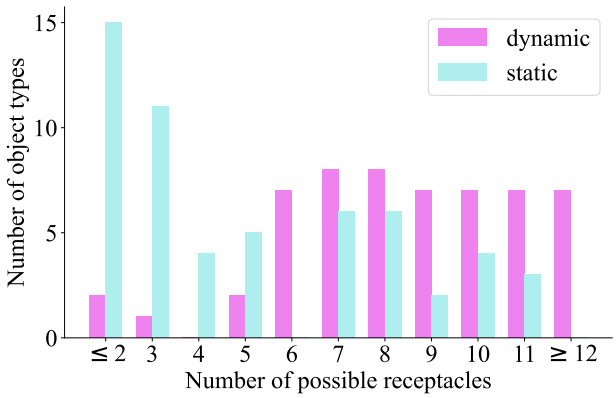


Fig. 4: Number of object types with different numbers of potential receptacle types.

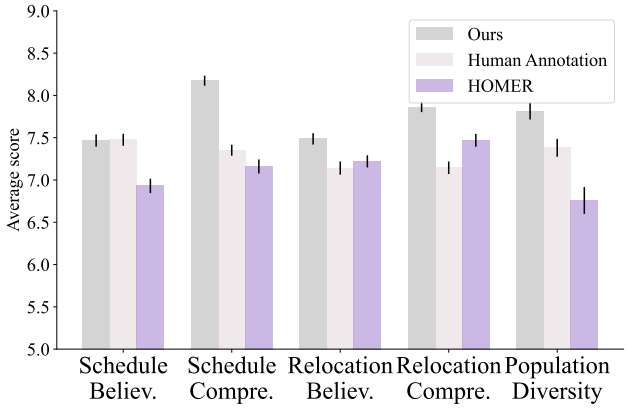


Fig. 5: Average scores in the user study, where the error bar reflects the standard error.

If there are multiple instances of the target receptacles in the sampled room, we randomly pick one.

V. Evaluation of the Generation Quality

A. Dataset Statistics

Our dataset of simulated human residents includes 50 distinct personas, with an average age of 35.8 ± 12.13 (mean and standard deviation, same hereinafter). We generate a 5-day schedule for each persona, from Wednesday to Sunday, costing approximately 60k tokens for each persona. Our dataset includes 564 activities, of which 448 activities are considered primary after merging synonyms within the schedules of each persona. Synonymous activities share the same environmental effects. In this study, we select the ProcTHOR dataset [11] as the source of static scenes. Visualization of the generated content with statistics of our dataset is presented in Figure 3. Aligning to the scene prior information, an event potentially involves 5.50 ± 4.58 objects, where the distribution is illustrated in Figure 3 (c). Compared to the static scenes, the placement of objects can also be more diversified in the dynamic ones, reflecting in the possible receptacles types of object types as shown in Figure 4.

B. User Study Designation

Since our dataset is largely based on synthetic data, it naturally requires an assessment of the quality of the human simulations to ensure that the dataset can support further research on dynamic goal navigation. To this end, we evaluate the generated human activities schedule and corresponding object relocations configured in dynamic scenes within the following dimensions:

- Believability, an idea borrowed from prior works on language agents [15, 42], indicating whether the generated content is reasonable and true to humans.
- Comprehensiveness, measuring whether the content is detailed enough to depict human behaviors without omitting important activities or object usage.
- Diversity. The generated content is expected to be diversified and personalized, instead of monotonously repeating the same routine.

Both believability and comprehensiveness are measured for both the schedules and the object relocations at the persona level, whereas diversity is measured at the method level, across multiple personas.

To verify the effectiveness of our framework, we compare it with the following baselines:

- HOMER dataset [2]. Although it captures delicate human data in a vivid simulation, the dataset is not directly applicable to the THOR simulators. We extract the human schedules and extract the corresponding object relocations from the environment variation for comparison at the text level.
- Human annotation. Following the conventional process of human annotation as adopted by [15], we recruit 10 college students (4 female, 6 male) to manually design a two-day schedule and corresponding object relocations for 15 distinct personas within 5 sampled scenes (3 personas per scene), with the same information as the input of our framework. The schedules are reviewed by the authors to ensure the quality.

We then conduct a user study, where 200 questionnaires are distributed through an international survey platform, each including the human behaviors acquired by all 3 methods in a single scene with 3 personas. We display the scene layouts including rooms, receptacles, and movable objects, along with the characteristics of personas, the generated human activity schedules, and the corresponding object relocations. For data from the HOMER dataset, we explicitly state that the activities are in a different scene and the persona is not specified. All subjects are ensured to be able to understand English by an embedded reading comprehension test. After filtering out those who failed to pass the reading comprehension test or failed to meet the minimum answer time requirement, a total of 166 questionnaires are collected for analysis.

C. Results

As the quantitative results presented in Figure 5, the simulated human behaviors and the corresponding object relocation achieve the highest overall assessment. Our generated schedules show a comparable level of believability to human annotations, while significantly outperforming the HOMER, probably due to the limited choice of activity in rule-based schedule synthesis methods. Benefiting from the hierarchical architecture, our framework can generate more detailed activities than the baselines. Surprisingly, our method also outperforms human annotation, we conjecture that humans may not be motivated enough to improve the detailedness of annotations. Not surprisingly, our framework exhibits better quality of object relocations than humans, since it might be hard to recognize all reasonable relocations among dozens of objects for human annotators. To our supervise, although the HOMER dataset models human behaviors more delicately, users tend to give higher appraisal to our framework, possibly being influenced by the better quality of schedules. Benefiting from the generation capabilities of LLMs, our framework can produce more diversified schedules based on multifarious personas, which are not considered by the HOMER dataset.

Overall, the results verify the quality concerning believability, comprehensiveness, and diversity of our dataset, which is comparable to or above average human level, further supporting the appropriateness of our dataset as a benchmark for studying dynamic semantic goal navigation.

VI. Navigation Experiments

A. Task and Baselines

To validate our generated scenes as potential embodied navigation benchmarks, we study the representative object goal navigation tasks as a demonstration. Our experiments cover various problem settings, including both map-free and map-based navigation. Considering the characteristics of dynamic scenes, we additionally introduce a human hint for each task, which is generated by asking the simulated humans whether they have recently used the target object. The hint is also generated by LLM by providing coarse-grained human activity schedules as human memory. To provide comprehensive results, we include various baselines that correspond to each problem setting.

a) Map-free Navigation.: Beyond the naive **random** agent that takes random actions, we evaluate the state-of-the art visual-based navigation models, EmbClip+Codebook and DINOv2+Codebook [25, 28], abbreviated as **EmbClip-C** and **DINOv2-C** hereinafter. We evaluate the publicly available checkpoints trained on the ProcTHOR dataset without further fine-tuning. Such a setting treats the dynamic scenes as *unseen* environments.

b) Map-based Navigation.: Considering the scenario that searching for movable daily objects in everyday life, the object-goal navigation in *seen* environments also worth studying, in which, with the aid of map constructing techniques [43, 44, 45], the problem can be simplified to map-based navigation, where the reachable points and positions of immovable receptacles are assumed to be known. The problem then turns into planning a route with estimated object distribution probabilities [5]. We first employ the one-step greedy search as a baseline (abbreviated as **OSG**), which is a computationally efficient approximation of CP-SAT introduced by [5]. We also propose a simple baseline that chooses the next exploring point according to the ratio of the estimated probability of finding the object to the navigation distance, namely cost-effective greedy (**CEG**).

c) Object Distribution Estimation for Map-based Navigation.: Planning-based algorithms require the estimation of probabilities to find the target on each receptacle, in this context, we include three baselines in the experiment: (1) Uniform, where all receptacles are considered equally possible to find the target object, and both the aforementioned planning algorithms degenerate into simply examine the nearest receptacles, denoted as **greedy**. (2) Scene prior (**SP**), where we estimate the probabilities according to the occurrences of object relationships in the 10,000 train environments in the ProcTHOR dataset, as an approach of learning-based commonsense reasoning. (3) Using **LLM**[2] to analyze the generated human hints, hierarchically extracting the relevant event from the hints and estimate the object placement. Though the human hints may introduce a too strong assumption, experiments show that navigation remains challenging even in this setting.

[2]We use gpt-4o-2024-05-13 throughout the navigation experiment section.

TABLE I: Results for the baseline models. The ↑ and ↓ denote larger and smaller values are preferred, respectively. Please note that the results are not directly comparable to those reported by [28], since our experiments are conducted in more difficult scenes and with slight different settings. Similarly, the results in map-free are not directly comparable to map-based ones.

Method	Task Setting		Normal			Hard		
	Map	Hint	EL↓	SR(%)↑	SPL↑	EL↓	SR(%)↑	SPL↑
Random	✗	✗	468.7	10.14	2.90	483.5	5.02	1.3
EmbClip-C	✗	✗	311.8	28.57	12.02	342.8	26.77	10.9
DINOv2-C	✗	✗	368.8	40.95	10.69	402.2	38.58	5.7
Greedy	✓	✗	161.5	41.30	16.02	176.49	22.69	10.14
SP + OSG	✓	✗	139.5	58.94	27.73	165.09	32.53	16.22
SP + CEG	✓	✗	129.2	68.12	34.56	159.51	36.35	20.57
LLM + OSG	✓	✓	140.7	59.42	27.33	165.25	31.53	16.14
LLM + CEG	✓	✓	107.4	76.81	46.13	146.65	46.59	28.56

All baselines are evaluated in zero-shot generalization settings, without fine-tuning on any dynamic scenes, while our experiment acquiesces to using the 10k static scenes in the training set of ProcTHOR.

B. Experimental Setup

We generate 100 scenes as the test set based on the test environments in the ProcTHOR dataset and our generated human activities, where half of the scenes are marked as **normal** difficulty (4-7 rooms with 1-2 simulated residents) and the others are regarded as **hard** ones (8-10 rooms with 3-4 simulated residents). For each scene, we sample up to 10 tasks at two instants of the dynamic scene, with 11 object types as potential targets. Only recently used objects will be selected as the target, resulting in a total of 414 tasks in the normal split and 498 tasks in the hard split, For all tasks, there exists only one instance of the target object.

Map-free methods are evaluated in similar settings to previous object goal navigation studies, where the agent is allowed to run 500 steps. However, since the object types in our dataset are not fully covered by [28], the visual-based models are only evaluated on 232 tasks, which is a subset of our benchmark. For map-based methods, we reduce the number to 200. The hyperparameter α_p in OSG is set to 0.5. In consideration of planning-based methods without a stop action, the task is considered successful if the distance between the agent and the target object is less than 1.5m and the target is visible to the agent. We adopted three popular navigation metrics: episode length (**EL**), success rate (**SR**), and success rate weighted by path length (**SPL**).

C. Results

We present the quantitative results in Table I. In the map-free setting, visual-based navigation models significantly outperform the random baseline, however, there still exists a large space for improvement. It is noteworthy that the scenes we used for evaluation are more difficult than the ProcTHOR dataset in various ways, potentially suggesting further aspects of improving the agents. In the map-based setting, we find both probability estimation and planning are essential. All evaluated methods significantly outperform the greedy baseline. We also find that the CEG planner outperforms the OSG, possibly due to our insufficient search of the hyperparameter. Moreover, we find that even with human hints, LLM + OSG fails to outperform SP + OSG, whereas the LLM brings a significant boost to the CEG, indicating that the quality of probability estimation may still require a strong planner to reflect in the quantitative metrics.

Generally, the results show that dynamic scenes introduce new challenges and task settings to object goal navigation. Nevertheless, we would like to emphasize that the results principally serve as a baseline and a demonstration of utilizing dynamic scenes as navigation benchmarks. We believe both the benchmark designation and navigation algorithms still have the potential to be unleashed in future works.

VII. Conclusions and Future Work

In this paper, we introduce a framework for simulating human activities and analyzing corresponding object relocation probabilities with LLMs thereby efficiently generating dynamic scenes based on static environments. A user study of the generated dynamic scenes validates the quality and believability of the simulated human activities. Furthermore, the navigation experiments exhibit the potential of the generated dynamic scenes serving as a benchmark for embodied navigation tasks, which introduce new challenges and task settings. We also demonstrate the usability of the dynamic scenes by conducting experiments with several representative baselines in various task settings.

Generally, our framework can serve as an infrastructure for dynamic scene generation. Nevertheless, our framework currently focuses on object relocations in relevantly large time scales, mostly at the hour level, and thus can be further improved in the fineness of simulation. Besides, this paper only performs the dynamic scene generation in the household domain based on the ProcTHOR dataset, leaving applications on more general domains as future work. Moreover, the navigation benchmark and experiments conducted remain preliminary, whereas the dynamic scenes include much more underlying information being unused, thus there may exist a lot of room for improvement for both the task designation and the model development.

References

[1] Yilun Du, Tomas Lozano-Perez, and Leslie Pack Kaelbling. Learning object-based state estimators for household robots. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12558–12565. IEEE, 2022.

[2] Maithili Patel and Sonia Chernova. Proactive robot assistance via spatio-temporal object modeling. In *6th Annual Conference on Robot Learning*, 2022.

[3] Maithili Patel, Aswin Gururaj Prakash, and Sonia Chernova. Predicting routine object usage for proactive robot assistance. In *Conference on Robot Learning*, pages 1068–1083. PMLR, 2023.

[4] Andrey Kurenkov, Michael Lingelbach, Tanmay Agarwal, Emily Jin, Chengshu Li, Ruohan Zhang, Li Fei-Fei, Jiajun Wu, Silvio Savarese, and Roberto Martın-Martın. Modeling dynamic environments with scene graph memory. In *International Conference on Machine Learning*, pages 17976–17993. PMLR, 2023.

[5] Sohan Rudra, Saksham Goel, Anirban Santara, Claudio Gentile, Laurent Perron, Fei Xia, Vikas Sindhwani, Carolina Parada, and Gaurav Aggarwal. A contextual bandit approach for learning to plan in environments with probabilistic goal configurations. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5645–5652. IEEE, 2023.

[6] Ji Ma, Hongming Dai, Yao Mu, Pengying Wu, Hao Wang, Xiaowei Chi, Yang Fei, Shanghang Zhang, and Chang Liu. Doze: A dataset for open-vocabulary zero-shot object navigation in dynamic environments. *arXiv preprint arXiv:2402.19007*, 2024.

[7] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

[8] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *2017 International Conference on 3D Vision (3DV)*, pages 667–676. IEEE, 2017.

[9] Haoyuan Fu, Wenqiang Xu, Ruolin Ye, Han Xue, Zhenjun Yu, Tutian Tang, Yutong Li, Wenxin Du, Jieyi Zhang, and Cewu Lu. Demonstrating RFUniverse: A Multiphysics Simulation Platform for Embodied AI. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023. doi: 10.15607/RSS.2023.XIX.087.

[10] Bokui Shen, Fei Xia, Chengshu Li, Roberto Martín-Martín, Linxi Fan, Guanzhi Wang, Claudia Pérez-D'Arpino, Shyamal Buch, Sanjana Srivastava, Lyne Tchapmi, et al. igibson 1.0: A simulation environment for interactive tasks in large realistic scenes. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7520–7527. IEEE, 2021.

[11] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. In *Advances in Neural Information Processing Systems*, 2022.

[12] Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, et al. Infinite photorealistic worlds using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12630–12641, 2023.

[13] Zun Wang, Jialu Li, Yicong Hong, Yi Wang, Qi Wu, Mohit Bansal, Stephen Gould, Hao Tan, and Yu Qiao. Scaling data generation in vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12009–12020, 2023.

[14] Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR, 2023.

[15] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, 2023.

[16] Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, et al. Robothor: An open simulation-to-real embodied ai platform. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3164–3174, 2020.

[17] Kiana Ehsani, Winson Han, Alvaro Herrasti, Eli VanderBilt, Luca Weihs, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Manipulathor: A framework for visual object manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4497–4506, 2021.

[18] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347, 2019.

[19] Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John M Turner, Noah D Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems*, 2021.

[20] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew

Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[21] Guangyao Zhai, Evin Pinar Örnek, Shun-Cheng Wu, Yan Di, Federico Tombari, Nassir Navab, and Benjamin Busam. Commonscenes: Generating commonsense 3d indoor scenes with scene graph diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[22] Xinpeng Wang, Chandan Yeshwanth, and Matthias Nießner. Sceneformer: Indoor scene generation with transformers. In *2021 International Conference on 3D Vision (3DV)*, pages 106–115. IEEE, 2021.

[23] Yandan Yang, Baoxiong Jia, Peiyuan Zhi, and Siyuan Huang. Physcene: Physically interactable 3d scene synthesis for embodied ai. *arXiv preprint arXiv:2404.09465*, 2024.

[24] Devendra Singh Chaplot, Dhiraj Gandhi, Abhinav Gupta, and Ruslan Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 4247–4258, 2020.

[25] Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. Simple but effective: Clip embeddings for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14829–14838, 2022.

[26] Haokuan Luo, Albert Yue, Zhang-Wei Hong, and Pulkit Agrawal. Stubborn: A strong baseline for indoor object navigation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3287–3293. IEEE, 2022.

[27] Albert J Zhai and Shenlong Wang. Peanut: predicting and navigating to unseen targets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10926–10935, 2023.

[28] Ainaz Eftekhar, Kuo-Hao Zeng, Jiafei Duan, Ali Farhadi, Aniruddha Kembhavi, and Ranjay Krishna. Selective visual representations improve convergence and generalization for embodied ai. In *The Twelfth International Conference on Learning Representations*, 2024.

[29] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8494–8502, 2018.

[30] Ifrah Idrees, Siddharth Singh, Kerui Xu, and Dylan F Glas. A framework for realistic simulation of daily human activity. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 30–37. IEEE, 2023.

[31] Ashish Mishra, Gyanaranjan Nayak, Suparna Bhattacharya, Tarun Kumar, Arpit Shah, and Martin Foltin. Llm-guided counterfactual data generation for fairer ai. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1538–1545, 2024.

[32] Stefan Sylvius Wagner, Maike Behrendt, Marc Ziegele, and Stefan Harmeling. Sqbc: Active learning using llm-generated synthetic data for stance detection in online political discussions. *arXiv preprint arXiv:2404.08078*, 2024.

[33] Yelaman Abdullin, Diego Molla-Aliod, Bahadorreza Ofoghi, John Yearwood, and Qingyang Li. Synthetic dialogue dataset generation using llm agents. *arXiv preprint arXiv:2401.17461*, 2024.

[34] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

[35] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*, 2023.

[36] Chenxu Wang, Bin Dai, Huaping Liu, and Baoyuan Wang. Towards objectively benchmarking social intelligence for language agents at action level. *arXiv preprint arXiv:2404.05337*, 2024.

[37] Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624): 1067–1074, 2022.

[38] Yuan Li, Yixuan Zhang, and Lichao Sun. Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents. *arXiv preprint arXiv:2310.06500*, 2023.

[39] Chuyi Kong, Yaxin Fan, Xiang Wan, Feng Jiang, and Benyou Wang. Platolm: Teaching llms via a socratic questioning user simulator, 2023.

[40] Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623(7987): 493–498, 2023.

[41] Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*, 2023.

[42] Joseph Bates et al. The role of emotion in believable agents. *Communications of the ACM*, 37(7):122–125, 1994.

[43] Gamini Dissanayake, Shoudong Huang, Zhan Wang, and Ravindra Ranasinghe. A review of recent developments in simultaneous localization and mapping. In *2011 6th International Conference on Industrial and Information Systems*, pages 477–482. IEEE, 2011.

[44] Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta,

Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural slam. In *International Conference on Learning Representations*, 2019.

[45] Devendra Singh Chaplot, Ruslan Salakhutdinov, Abhinav Gupta, and Saurabh Gupta. Neural topological slam for visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12875–12884, 2020.