Joint Human-Scene Generation with Mutual Spatial and Functional Awareness and Flexible Control

Supervisor/Contact: Chenliang Zhou (chenliang.zhou@cst.cam.ac.uk)

Introduction

The generation of 3D scenes has garnered significant attention due to its wide-ranging applications in industries such as gaming, architecture, virtual environments, and synthetic data creation for AI and machine learning. Existing research has explored *either* **human-aware scene generation** – where human movements are influenced by a pre-generated 3D scene (e.g., MIME [1]), ensuring, for example, that objects do not overlap with human poses (i.e., **spatial awareness**), and **scene-aware human motion generation**, where human motion adapts to the given 3D environment (e.g., TeSMo [9]), avoiding obstacles, and possibly interacting with objects (i.e., **functional awareness**) like sitting in chairs, laying on beds, or opening fridges. However, few approaches consider **joint generation of humans and objects** in a scene with **mutual spatial and functional awareness**.

Many real-world applications require a more integrated approach: the simultaneous generation of both 3D scenes and human motion, with both being mutually aware of each other's presence and constraints. For instance, in simulations for gaming or virtual reality, the placement of objects and human movements within the same scene should be dynamically coordinated for better realism and interactivity. By introducing **mutual awareness**, we aim to create a unified pipeline where the 3D object and human motion generation are informed by one another, providing more coherent and contextually appropriate outputs. This approach would be particularly useful in:

- Enhancing virtual simulation environments where human-object interactions are key.
- Applications requiring spatial realism, such as urban planning, emergency response simulations, and entertainment (e.g., VR, gaming).
- Synthetic data generation for training AI models in human-environment interactions.

This project proposes a unified framework that **simultaneously generates both humans and objects** within a 3D scene, ensuring coherence, non-overlap, and realistic human-scene interaction. The approach builds on and integrates existing modules from the two sub-tasks (human-conditioned scene generation and scene-conditioned human generation), offering a practical and scalable starting point.

To extend flexibility, we also explore **multi-modal control signals** for conditional generation, including:

- **Text prompts** describing the scene or activities
- **Floorplan-based maps** specifying where humans or objects should appear

(optionally with semantic types)

- **Room layout sketches or 2D images** as structural constraints

This joint and controllable generation paradigm can unlock more realistic, interactive, and user-guided scene synthesis, with potential applications in AR/VR, robotics simulation, and digital content creation.

Goal

The objective of this project is to develop a unified generative framework that jointly synthesizes 3D scenes with both humans and objects, ensuring mutual spatial and functional awareness, and enabling flexible, user-controllable generation via multi-modal inputs (e.g., text, floor maps, or sketches).

Task

1 **Literature review**: Familiarize yourself with key works related to 3D scene generation (e.g., [1-6]) and human motion generation (e.g., [7-9]). Special attention should be given to models such as MIME [1] (human-aware scene generation) and TeSMo [9] (scene-aware human motion generation). Essentially, our goal is to create a unified pipeline that combines these capabilities. While the listed works provide a good foundation, you are encouraged to explore additional related literature (e.g., using Google Scholar or ArXiv) to identify cutting-edge approaches in the field.

2 **Identify and understand a baseline model**: Identify the most relevant study that will serve as the foundation for our pipeline. Study this model in detail and ensure you are able to successfully run its codebase. If applicable, reuse their rendering pipeline to visualize the generated 3D scene and human motion, which will save time and ensure compatibility. Ensure you can generate visualized results early on to expedite future evaluations.

3 **Data collection or generation**: Identify or generate a suitable dataset that includes both **human motions** and **3D scenes**. This could involve combining existing datasets or even creating a new one by simulating environments and recording human interactions. Ensure the dataset includes diverse scene layouts and a variety of human actions to enrich the model's learning process. You can start from the ones used in [1] and/or [9].

4 **Design and develop the pipeline**: Discuss and decide on the architecture of the pipeline. Will it be based on **GANs, diffusion models**, or some other generative method? Will the scene and human motions be generated simultaneously, or will there be sequential generation steps? Consider how mutual awareness will be implemented—should there be a feedback loop between scene layout and human movement during generation? Is there anything we can learn from the existing studies?

5 **Performance evaluation**: Evaluate the performance of the pipeline using both **qualitative** and **quantitative** metrics. For qualitative evaluation, visualize

the coherence between the scene and human motion. Quantitative metrics could include measures of spatial realism, scene occupancy, or plausibility of human motion. Additionally, user studies or expert evaluation could be employed to assess the model's outputs.

6 **Possible extensions**:

6.1 **Spatially controlled generation**: Enhance the user's control over the generated scene by enabling spatial constraints. For example, allow the user to specify where specific types of objects or humans should (or should not) appear. This can be achieved by conditioning the model on a **floor plan** (used in [1]) or spatial map.

6.2 **Text-guided generation**: Incorporate **text-based guidance** to generate scenes and motions based on user-provided descriptions. This can be informed by recent advancements in scene [5] and human motion generation [7-9] conditioned on natural language prompts.

6.3 **Generation from 2D images**: Develop functionality to generate 3D scenes and human motion from 2D images of rooms, furniture layouts, or human activities. This could be achieved through the integration of 2D-to-3D generation techniques, extending the application of your model to areas such as interior design or object layout prediction.

6.4 **Use of diffusion models**: Investigate the use of **diffusion models** [4, 5, 6, 9] for the simultaneous generation of 3D scenes and human motions. Diffusion models have shown remarkable success in both image and motion generation tasks, and leveraging these advancements could improve the quality and realism of the generated outputs.

References

[1] https://mime.is.tue.mpg.de
[2] https://github.com/musialski-research/LayoutEnhancer
[3] https://xinpeng-wang.github.io/sceneformer/
[4] https://tangjiapeng.github.io/projects/DiffuScene/
[5] https://lang-scene-synth.github.io
[6] https://physcene.github.io
[7] https://neu-vi.github.io/omnicontrol/
[8] https://tlcontrol.weilinwl.com
[9] https://research.nvidia.com/labs/toronto-ai/tesmo/