# CLIPGraphs: Multimodal graph networks to infer object-room affinities for scene rearrangement

Ayush Agrawal*[1], Raghav Arora*[1], Ahana Datta[1], Snehasis Banerjee[1,2], Brojeshwar Bhowmick[2], Krishna Murthy Jatavallabhula[4], Mohan Sridharan[3], Madhava Krishna[1]

[1]Robotics Research Center, IIIT Hyderabad, India
[2]TCS Research, Tata Consultancy Services, India
[3]Intelligent Robotics Lab, University of Birmingham, UK
[4]Massachusetts Institute of Technology, USA

***Abstract*—Computing the most suitable room for any given object is an important step in the scene rearrangement challenge for embodied AI. State of the art methods use LLMs or reinforcement learning-based policies for this task. We propose CLIPGraphs, a method that leverages the complementary strengths of reasoning with commonsense knowledge and data-driven methods for this task. Specifically, it encodes a knowledge graph of prior human annotations of the occurrence of different objects in particular rooms in home environments, incorporates features from vision-language models to support multimodal queries based on images or text, and uses a graph neural network to learns object-room affinities based on embeddings of prior knowledge and the vision-language model features. We demonstrate that our approach provide better estimates of the most appropriate location of objects from a benchmark set of categories in comparison with state of the art baselines.** [1]

***Index Terms*—Commonsense knowledge, graph convolutional network, knowledge graph, large language models**

*Denotes equal contribution
[1]Supplementary material and code: Link