Proceedings of the Eighteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment
(AIIDE 2022)

Explainable CLIP-Guided 3D-Scene Generation in an AI Holodeck

Atefeh Mahdavi Goloujeh, Jason Smith, Brian Magerko

Georgia Institute of Technology
85 5th St NW Atlanta, Georgia 30308
{atefehmahdavi, jsmith775, magerko}@gatech.edu

Abstract

This paper describes the *AI Holodeck*, a co-creative software prototype that creates virtual scenes from the user input text, inspired by the fictional Holodeck virtual reality device from the science fiction series *Star Trek*. This application collects common-sense knowledge from annotated datasets and reference images. It uses this knowledge to populate scenes with objects found in selected environments alongside those explicitly mentioned by the user. We present the system design of the *AI Holodeck*, and a proposed study to measure the effects of its visualizations on user perceptions of the system's creativity.

Introduction

One major challenge of text-to-scene generation is the generation of scenes that are diverse yet maintain their relevance to the user input. Prior work on 3D scene generation has focused largely on the plausibility of the scenes (Chang et al. 2015; Coyne and Sproat 2001) using objects that are explicitly mentioned by the user or relevant objects without considering the base environment. However, virtual scenes can also contain *implicit objects* (i.e. objects that are related to others through common sense and derived from implicit scene knowledge (Öhlschläger and Võ 2020)). Implicit objects can be gathered through either environment-specific or instance-specific knowledge and can improve the plausibility of a generated scene by portraying the realism of a diverse and densely-populated physical space.

We have previously presented the initial phases of the *AI Holodeck* (Smith et al. 2021), an application to generate virtual 3D scenes from natural language input using environment-specific knowledge we gathered through an annotated dataset. In this paper, we present a new iteration of the *AI Holodeck* application that gathers common-sense knowledge through two layers. First, as seen in the previous version, an annotated image dataset provides the system with objects implicitly related to those explicitly defined by the user. Second, a CLIP-guided (Radford et al. 2021) search extracts objects and their spatial relationships from reference images related to the user input. This addition also increased the need for explainability in our system due to the presence of surprising implicit objects or their placement. For example, Unexpected instance-specific objects may be the outcome of a particular uncommon instance or our system's failure to detect particular objects and their positional relationships. In the *AI Holodeck*, we discuss explainability as an approach to explicitly show whether something is out of place due to an error or simply an unexpected instance-specific object.

Related Work

Generating images from text has gained a large amount of attention in applications developed over the past few years (Galatolo., Cimino., and Vaglini 2021; Gafni et al. 2022; Wang et al. 2022; Li et al. 2019). Many of these applications leverage a pre-trained model CLIP (Contrastive Language–Image Pretraining) that captures semantics visual concepts from images (Radford et al. 2021). A notable recent AI system *Dall-E* can generate significant images or make edits to images from natural language using CLIP (Ramesh et al. 2022). The outputs of *Dall-E* and similar systems are significant in the context of 2D image generation. However, transferring these advancements to the context of 3D scene generation is yet to be explored.

There have been several methods exploring text-to-3D shape and mesh generation using CLIP (Khalid et al. 2022; Hong et al. 2022; Canfes et al. 2022). For example, a method called CLIP-Forge leverages CLIP to present a zero-shot text-to-shape generation system (Sanghi et al. 2021). While these works explore 3D shapes, our work focuses on finding relevant objects and their arrangement in 3D space.

System Overview

The *AI Holodeck* application creates full virtual scenes from user text input (Smith et al. 2021). It parses user input into a dependency tree consisting of subjects, objects, and descriptors using either the CoreNLP (Manning et al. 2014) or NLTK (Bird, Klein, and Loper 2009) library. Each object is stored as a node in a scene template that keeps track of different objects and their positional relationships.

The first iteration of the *AI Holodeck* analyzed images to fill scenes with appropriate implicit objects alongside those explicitly mentioned by the user (Smith et al. 2021). This new iteration uses that technique alongside a combination of existing CLIP search and Graph R-CNN (Yang et al. 2018)

methods to find objects and positional relationships in an image based on user input.

Implicit Object Selection

The MIT Indoor Scenes dataset (Quattoni and Torralba 2009) contains a series of 2D images alongside the boundaries of each object in the image. We developed a rule-based architecture to estimate the positional relationships of objects in relation to each other based on the provided boundaries (Smith et al. 2021). This process generates a dictionary of objects and 4 directions (above, below, left, and right) for each. By aggregating the entire dataset, we create a dictionary of objects and their probability of occurrence with other objects in these cardinal directions. This dictionary is then used to add implicit Objects to a scene.

Explainable CLIP-Guided Search

The main addition to the *AI Holodeck* described in this paper is a second method of adding implicit objects to scenes and converting their 2D representations to 3D, which combines a CLIP search with Graph R-CNN (Yang et al. 2018). Since images beyond the MIT indoor dataset do not come with annotations and boundaries, we used the pre-trained Graph R-CNN to annotate images searched through CLIP. The pre-trained Graph R-CNN detects objects and predicts positional relationships but does not output a complete representation of a 3D environment. The output of this model for each image is a list of pairs of objects and their positional relationship. For example, the output of the model for an image can be (vase, desk, top), showing that the model detected a vase and a desk, and the vase is on top of desk. The *AI Holodeck* application uses the user's input text as input to CLIP search, finds and annotates an appropriate image, and combines the objects and relationships from the image with the ones found through the initial MIT Indoor Scenes dataset method (Figure 1). The found annotated image is also presented to users as a visual explanation (Figure 2).

Scene Visualization

The generated objects and relationships are added to the *AI Holodeck*'s internal Scene Template. By referencing the objects' relationships and dimensions in the ShapeNet database, the system creates a 3D bounding box for each object and placing subsequent objects in their cardinal directions. The visualization algorithm begins with objects at the bottom of the scene and recursively adds objects.

Discussion and Future Work

A lack of a formal evaluation is a current limitation of the work. We have a study planned to determine the effects of visual explanation in the form of the pictures accompanying the CLIP search, on how users perceive the relevance of scenes generated by the *AI Holodeck* to user input, and in the creativity exhibited by the system. Participants will answer surveys before and after a guided scene generation task– with and without a display of the visual explanation.

Participants in our planned evaluation will be tasked with thinking of an environment they want to make, describing

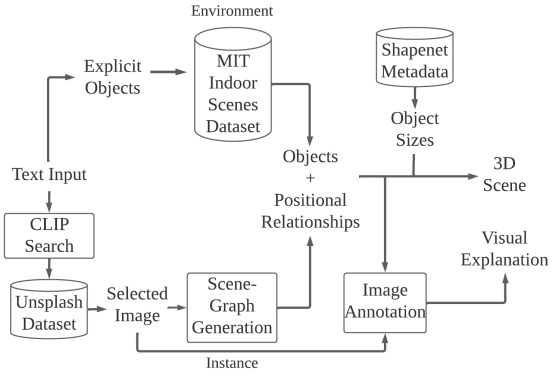

Figure 1: Pipeline for insertion of implicit objects.

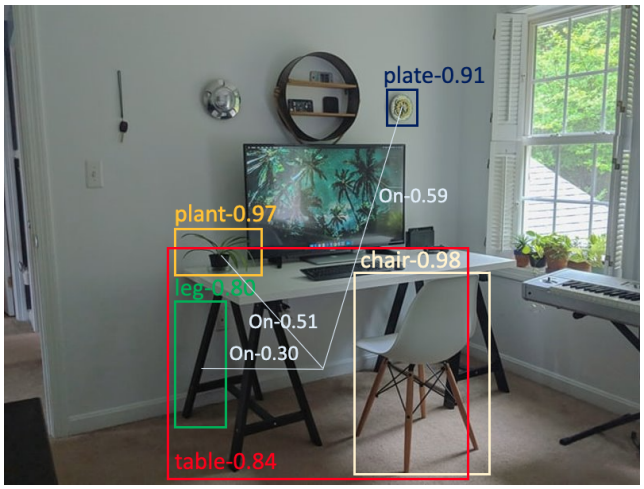

Figure 2: An example annotated visual explanation

it, and then adding multiple sentences of text input to the *AI Holodeck* application. All subjects will be using one version without any displayed CLIP search image. Half of the subjects will also use a version with a CLIP image, and the others will use a version with the annotated CLIP search image. After using each version, subjects will measure their perceptions of each version's ability to support them creatively (Cherry and Latulipe 2014) and will compare their satisfaction and trust in the explanations generated by each version (Hoffman et al. 2018).

The *AI Holodeck* application provides a case study of leveraging "common sense" positional knowledge to generate creative scenes. By evaluating how the explainability afforded by annotated visualizations affects user perceptions of creativity in this text-to-scene generation application, we will contribute to the Explainable AI community by how visualizing a single component of a multi-model system affects user perception of that system's creativity. These future findings could ideally also be applied to other creative or open-ended domains such as music composition, narrative generation, or game design.

References

Bird, S.; Klein, E.; and Loper, E. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Canfes, Z.; Atasoy, M. F.; Dirik, A.; and Yanardag, P. 2022. Text and Image Guided 3D Avatar Generation and Manipulation. *arXiv preprint arXiv:2202.06079*.

Chang, A.; Monroe, W.; Savva, M.; Potts, C.; and Manning, C. D. 2015. Text to 3d scene generation with rich lexical grounding. *arXiv preprint arXiv:1505.06289*.

Cherry, E.; and Latulipe, C. 2014. Quantifying the creativity support of digital tools through the creativity support index. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 21(4): 1–25.

Coyne, B.; and Sproat, R. 2001. WordsEye: An automatic text-to-scene conversion system. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 487–496.

Gafni, O.; Polyak, A.; Ashual, O.; Sheynin, S.; Parikh, D.; and Taigman, Y. 2022. Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors. *arXiv preprint arXiv:2203.13131*.

Galatolo., F.; Cimino., M.; and Vaglini, G. 2021. Generating Images from Caption and Vice Versa via CLIP-Guided Generative Latent Space Search. *Proceedings of the International Conference on Image Processing and Vision Engineering*.

Hoffman, R. R.; Mueller, S. T.; Klein, G.; and Litman, J. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

Hong, F.; Zhang, M.; Pan, L.; Cai, Z.; Yang, L.; and Liu, Z. 2022. AvatarCLIP: Zero-Shot Text-Driven Generation and Animation of 3D Avatars. *arXiv preprint arXiv:2205.08535*.

Khalid, N.; Xie, T.; Belilovsky, E.; and Popa, T. 2022. Text to Mesh Without 3D Supervision Using Limit Subdivision. *arXiv preprint arXiv:2203.13333*.

Li, B.; Qi, X.; Lukasiewicz, T.; and Torr, P. 2019. Controllable text-to-image generation. *Advances in Neural Information Processing Systems*, 32.

Manning, C. D.; Surdeanu, M.; Bauer, J.; Finkel, J. R.; Bethard, S.; and McClosky, D. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 55–60.

Öhlschläger, S.; and Võ, M. L.-H. 2020. Development of scene knowledge: Evidence from explicit and implicit scene knowledge measures. *Journal of Experimental Child Psychology*, 194: 104782.

Quattoni, A.; and Torralba, A. 2009. Recognizing indoor scenes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 413–420. IEEE.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Sanghi, A.; Chu, H.; Lambourne, J. G.; Wang, Y.; Cheng, C.-Y.; and Fumero, M. 2021. Clip-forge: Towards zero-shot text-to-shape generation. *arXiv preprint arXiv:2110.02624*.

Smith, J.; Anaraki, N. A. T.; Goloujeh, A. M.; Khosla, K.; and Magerko, B. 2021. Towards an AI Holodeck: Generating Virtual Scenes from Sparse Natural Language Input.

Wang, Z.; Liu, W.; He, Q.; Wu, X.; and Yi, Z. 2022. CLIP-GEN: Language-Free Training of a Text-to-Image Generator with CLIP. *arXiv preprint arXiv:2203.00386*.

Yang, J.; Lu, J.; Lee, S.; Batra, D.; and Parikh, D. 2018. Graph r-cnn for scene graph generation. In *Proceedings of the European conference on computer vision (ECCV)*, 670–685.