Scene Understanding via Scene Representation Generation with Vision-Language Models

Yuan Chen, Peng Shi
School of Computer Science, University of Waterloo, Canada
constant.chen, peng.shi@uwaterloo.ca

Abstract

Understanding complex environments requires capturing the arrangement of objects, their interactions, and contextual information. Early symbolic and data-driven approaches are limited by rigid designs or narrow applicability. Recent vision-language models (VLMs) provide rich priors and flexible reasoning, supporting the generation of structured scene descriptions that handle compositional arrangements, diverse categories, and realistic constraints. However, challenges remain in precise spatial reasoning, consistent object placement, and maintaining coherent geometry. We present a VLM-driven pipeline for scene representation generation, analyze its shortcomings through a case study, and suggest avenues for future enhancements.

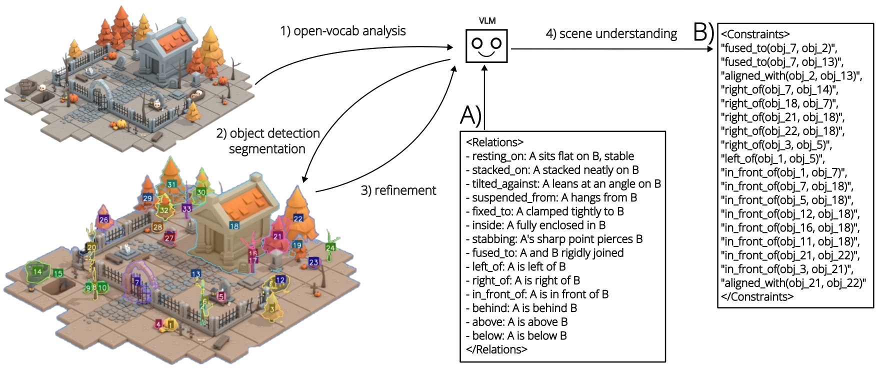
1 Introduction

Scene understanding is a fundamental problem in computer vision and robotics, involving the interpretation of spatial layouts, object relationships, and semantic context within complex environments. To support such understanding, scene representations provide structured, machine-interpretable descriptions of 2D or 3D environments. Early rule-based or template-based approaches [3, 4] require careful hand-crafted design for spatial understanding; existing data-driven methods for modelling spatial relationships [2, 11] are constrained by their narrow semantic scope and limited flexibility across different environments. Vision-language models (VLMs) relax these constraints and open new avenues by serving as priors that imbue systems with world knowledge and linguistic flexibility unattainable in earlier approaches. They plan, generate, and refine scene descriptions, addressing core challenges in scene understanding—including compositional complexity (arranging multiple objects with spatial relationships), semantic generalization (open-vocabulary understanding beyond fixed catalogues), physical realism (ensuring scenes are physically plausible), and controllability (enabling user guidance and iterative editing). However, VLMs still struggle with fine-grained spatial reasoning, consistent object grounding, and maintaining geometric coherence across complex scenes. Their representations often capture high-level semantics but fail to model precise spatial arrangements or physical constraints. In this work, we present a pipeline that leverages the VLMs for scene representation generation. Based on a case study, we identify current limitations and envision potential directions for future improvements.

2 VLMs for Generating Scene Representation

Recent advances in VLMs have enabled new approaches for generating and reasoning over structured scene representations. Two commonly used representation formats are programmatic and graph-based structures. **Programmatic formats** typically consist of code blocks written in languages such as Python or CSS, which encode information about object positions, rotations, and scales. For example, LayoutGPT [5] represents visual layouts using CSS-style code, which can be interpreted into 2D or 3D spatial compositions while supporting explicit spatial constraints. Similarly, SceneCraft [7] iteratively

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: .



```
<Relations>
- resting_on: A sits flat on B, stable
- stacked_on: A stacked neatly on B
- tilted_against: A leans at an angle on B
- suspended_from: A hangs from B
- fixed_to: A clamped tightly to B
- inside: A fully enclosed in B
- stabbing: A's sharp point pierces B
- fused_to: A and B rigidly joined
- left_of: A is left of B
- right_of: A is right of B
- in_front_of: A is in front of B
- behind: A is behind B
- above: A is above B
- below: A is below B
</Relations>
```

```
<Constraints>
"fused_to(obj_7, obj_2)",
"fused_to(obj_7, obj_13)",
"aligned_with(obj_2, obj_13)",
"right_of(obj_7, obj_14)",
"right_of(obj_18, obj_7)",
"right_of(obj_21, obj_18)",
"right_of(obj_22, obj_18)",
"right_of(obj_3, obj_5)",
"left_of(obj_1, obj_5)",
"in_front_of(obj_1, obj_7)",
"in_front_of(obj_7, obj_18)",
"in_front_of(obj_5, obj_18)",
"in_front_of(obj_12, obj_18)",
"in_front_of(obj_16, obj_18)",
"in_front_of(obj_11, obj_18)",
"in_front_of(obj_21, obj_22)",
"in_front_of(obj_3, obj_21)",
"aligned_with(obj_21, obj_22)"
</Constraints>
```

Figure 1: The pipeline of leveraging VLM for scene representation generation.

generates Blender Python scripts guided by visual feedback from VLMs to manipulate objects within the scene. On the other hand, **graph-based approaches** first assemble object relationships and constraints without specifying detailed position, rotation, or scale information. Precise object placement is then delegated to traditional solvers, enabling modular scene generation while ensuring physical validity [1, 9, 6, 8, 13].

We adopt a graph-based representation in our pipeline for greater semantic flexibility and generality in practice. Given an image, the pipeline begins with a VLM listing objects, tagged with open vocabularies. The resulting list is then passed to a detection and segmentation module. We further mark each detected object in the image to improve VLM's grounding performance. Then the VLMs are prompted to generate pairwise spatial relationships to construct a coherent scene representation.

3 Case Study

We used an off-the-shelf Halloween asset pack [1] for a case study. Based on the pipeline, we evaluated how VLMs actually perform at scene understanding in practice, as shown in Figure 1. We first prompt GPT-5 [2] to identify visible objects in the scene and produce an open-vocabulary category list. We then passed the list to Grounded-SAM module [10] to detect and segment instances in the scene. We further applied Set-of-Mark (SoM) [12] to assign unique labels to each segmented object.

We followed [8] to design a scene language (shown in Figure 1 (A)) to encode spatial relationships among objects. The parse results from VLM (GPT-5 in our experiment) are shown in (B). The parsed scene representation exposed several issues. Some of the relations were missed, such as $right\_of(obj_1, obj_7)$. The VLM also failed to account for clutter and occlusion. In particular, it did not reliably recognize partially occluded tree instances labelled 29–33.

These preliminary results highlight three practical challenges: (1) 3D spatial inconsistencies, arising from incomplete and locally inconsistent pairwise relations within the scene; (2) limitations in downstream reasoning for dense outdoor environments; and (3) incomplete open-vocabulary object discovery, where fine-grained categories are missed due to insufficient segmentation. These issues underscore the need for geometric validation to ensure globally consistent pairwise relations, as well as occlusion-aware instance recovery to enhance the robustness of VLMs in real-world scenes.

4 Conclusion

While VLMs have advanced scene understanding, they still face challenges in spatial reasoning under occlusion and scene clutter. Future work should aim to develop a more systematic scene-language framework that enforces geometric consistency, explicitly models occlusion, and enables robust downstream control in complex environments.

[1] https://kaylousberg.itch.io/halloween-bits, Licensed under (Creative Commons Zero, CC0)

[2] GPT-5 features state-of-the-art performance across coding, math, and visual perception.

References

[1] Ata Çelen, Guo Han, Konrad Schindler, Luc Van Gool, Iro Armeni, Anton Obukhov, and Xi Wang. I-design: Personalized llm interior designer. In *European Conference on Computer Vision*, pages 217–234. Springer, 2024.

[2] Angel Chang, Manolis Savva, and Christopher D Manning. Learning spatial knowledge for text to 3d scene generation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 2028–2038, 2014.

[3] Bob Coyne and Richard Sproat. Wordseye: an automatic text-to-scene conversion system. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '01, page 487–496, New York, NY, USA, 2001. Association for Computing Machinery.

[4] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*, pages 15–29. Springer, 2010.

[5] Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 36:18225–18250, 2023.

[6] Gege Gao, Weiyang Liu, Anpei Chen, Andreas Geiger, and Bernhard Schölkopf. Graphdreamer: Compositional 3d scene synthesis from scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21295–21304, 2024.

[7] Ziniu Hu, Ahmet Iscen, Aashi Jain, Thomas Kipf, Yisong Yue, David A Ross, Cordelia Schmid, and Alireza Fathi. Scenecraft: An llm agent for synthesizing 3d scenes as blender code. In *Forty-first International Conference on Machine Learning*, 2024.

[8] Lu Ling, Chen-Hsuan Lin, Tsung-Yi Lin, Yifan Ding, Yu Zeng, Yichen Sheng, Yunhao Ge, Ming-Yu Liu, Aniket Bera, and Zhaoshuo Li. Scenethesis: A language and vision agentic framework for 3d scene generation. *arXiv preprint arXiv:2505.02836*, 2025.

[9] Gabrielle Littlefair, Niladri Shekhar Dutt, and Niloy J Mitra. Flairgpt: Repurposing llms for interior designs. In *Computer Graphics Forum*, page e70036. Wiley Online Library, 2025.

[10] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

[11] Manolis Savva, Angel X Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. Pigraphs: learning interaction snapshots from observations. *ACM Transactions On Graphics (TOG)*, 35(4):1–12, 2016.

[12] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023.

[13] K Yao, L Zhang, X Yan, Y Zeng, Q Zhang, L Xu, W Yang, J Gu, and J Yu. Cast: Component-aligned 3d scene reconstruction from an rgb image. arxiv 2025. *arXiv preprint arXiv:2502.12894*, 2025.