

Annotated Bibliography

Leslie Osei-Anane

February 18, 2026

References

- [1] 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*, volume 35. Curran Associates, Inc., 2022.

ProcTHOR introduces a procedural generation framework for creating large numbers of diverse, fully interactive indoor environments for embodied AI. The system samples floorplans, populates them with a large library of interactive objects, and randomizes layouts, materials, and lighting to scale scene diversity. The paper contributes both the ProcTHOR generation pipeline and a complementary artist-designed evaluation set (ArchitecTHOR), and it reports that training on large procedurally generated collections improves performance and generalization across multiple embodied AI benchmarks. This is a strong scholarly source because it is a peer-reviewed NeurIPS 2022 contribution with extensive empirical evidence and a clear methodological advance in scalable environment creation, which is central to embodied AI research. It also clarifies trade-offs between procedural controllability and realism that matter when selecting generation settings. This connects to the Holodeck-like concept by providing a concrete example of deterministic compilation from structured scene specifications into simulated environments and by outlining evaluation protocols for those environments in embodied tasks.

- [2] Weixi Feng, Wanrong Zhu, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Xuehai He, S. Basu, Xin Eric Wang, and William Yang Wang. LayoutGPT: Compositional visual planning and generation with large language models. In *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., 2023.

LayoutGPT investigates large language models as visual planners that convert natural-language prompts into structured layouts, reducing the burden of manual, low-level scene specification. The paper proposes in-context demonstrations in a stylesheet-like representation so an LLM can infer spatial and numerical constraints and output coherent compositions. The authors evaluate across multiple domains, including 3D indoor layout

synthesis, and report competitive performance relative to supervised methods while maintaining high controllability for spatial relations. This is a strong scholarly source because it is a peer-reviewed NeurIPS 2023 contribution with broad experiments and direct baseline comparisons. It is relevant because it operationalizes an important separation of concerns: planning in symbolic/structured space first, then generation/rendering afterward. For the Holodeck architecture, LayoutGPT directly motivates the planner stage that maps prompt text into a schema-locked WorldSpec rather than raw geometry. Its results support using constrained intermediate outputs for reliability, explainability, and deterministic compilation, which are core requirements when building VR worlds that must satisfy walkability, safety, and runtime budget checks.

- [3] 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 (CVPR)*, pages 21295–21304. IEEE/CVF, June 2024.

GraphDreamer proposes generating compositional 3D scenes from scene graphs, where objects are nodes and relationships are edges. The method addresses a common limitation in text-only conditioning for complex scenes by preserving explicit relational structure through the generation pipeline. It leverages pretrained text-to-image priors while introducing object-wise disentanglement and relationship-aware constraints, including mechanisms to reduce object interpenetration. The paper also demonstrates a practical route from natural language to structured graph input, then to 3D scene synthesis. The source is credible because it is peer-reviewed at CVPR 2024 and presents a clear problem definition, method, and empirical validation. It is especially relevant to this project because scene-graph conditioning mirrors the role of a structured intermediate representation in deterministic compilation. For Holodeck, GraphDreamer supports the idea that prompt parsing should produce explicit, machine-checkable structure before rendering. Its compositional approach maps well to a WorldSpec-like contract: entities, relations, and constraints are represented directly, which improves controllability, debugging, and reproducibility compared with unconstrained end-to-end generation.

- [4] Lukas Höllerin, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2Room: Extracting Textured 3D Meshes from 2D Text to Image Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7909–7920. IEEE/CVF, October 2023.

Text2Room presents a pipeline that transforms a text prompt into a room-scale textured 3D mesh by iteratively lifting outputs from text-to-image models into a multi-view consistent 3D representation. The method combines view synthesis, depth-based geometry reasoning, and texture projection

to progressively build a coherent scene mesh, while enforcing cross-view consistency and reducing geometric drift. A major contribution is showing that strong 2D generative priors can be operationalized for 3D scene creation without requiring a fully trained end-to-end text-to-3D model. This is a strong scholarly source because it is a peer-reviewed ICCV 2023 paper with reproducible methodology, controlled comparisons, and detailed qualitative and quantitative analysis. It is highly relevant for prompt-to-world systems because it directly addresses the key engineering challenge of converting language-guided image priors into navigable 3D structure. For the Holodeck project, Text2Room informs the world-building side of the architecture: a planner can emit structured constraints while a generator reconstructs spatially consistent geometry from those constraints. Even if the MVP uses prefab composition, Text2Room provides a roadmap for future upgrades toward richer geometry synthesis while preserving deterministic compiler checks.

- [5] Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, Chris Callison-Burch, Mark Yatskar, Aniruddha Kembhavi, and Christopher Clark. Holodeck: Language Guided Generation of 3D Embodied AI Environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16227–16237. IEEE/CVF, June 2024.

Holodeck presents a system that converts a natural-language prompt into an interactive 3D environment for embodied AI. The approach uses a large language model to draft object inventories and spatial relations, then optimizes object placement to satisfy those constraints while populating scenes with assets from Objaverse. The paper contributes a full prompt-to-scene pipeline, human preference studies that outperform procedural baselines on residential scenes, and demonstrations that agents can be trained in the generated environments. This is a credible source because it is a peer-reviewed CVPR 2024 paper with a clear methodology, extensive evaluations, and transparent comparisons to baselines, which strengthens confidence in the reported gains. It is relevant to research on language-guided environment generation and controllable simulation. This connects to the Holodeck-like prompt-to-world/spec generation concept by showing how constraints can be explicitly represented, solved deterministically, and evaluated in simulated environments to assess downstream agent performance.

- [6] 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 *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., 2023.

CommonScenes presents a fully generative framework that produces controllable 3D indoor scenes from scene graphs while preserving commonsense consistency. The method combines a layout branch and a shape-generation

branch so both global arrangement and local object geometry are modeled jointly. By conditioning on scene-object and object-object relationships, the approach improves semantic coherence compared with retrieval-heavy alternatives. The paper also contributes SG-FRONT, a graph-augmented dataset built on 3D-FRONT, enabling more rigorous evaluation of relationship-aware scene generation. This is a credible source because it is peer-reviewed at NeurIPS 2023, includes clear methodological components, and evaluates consistency, diversity, and quality against prior methods. It is directly relevant to this project because controllability and relational consistency are central to dependable prompt-to-world systems. For Holodeck, CommonScenes reinforces the value of scene-graph or spec-based planning before compilation. Its evidence suggests that explicit relational conditioning improves scene validity, which aligns with deterministic validation goals such as walkability checks, object compatibility, and predictable regeneration from the same prompt and seed.

- [7] Songchun Zhang, Yibo Zhang, Quan Zheng, Rui Ma, Wei Hua, Hujun Bao, Weiwei Xu, and Changqing Zou. 3D-SceneDreamer: Text-driven 3D-consistent scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10170–10180. IEEE/CVF, June 2024.

3D-SceneDreamer targets text-driven scene synthesis with explicit emphasis on preserving 3D consistency across viewpoints. The paper identifies a core weakness in prior warping-and-inpainting pipelines: cumulative geometric and appearance errors. To address this, the authors use a tri-plane feature NeRF as a global scene representation and introduce a generative refinement stage that repeatedly aggregates global 3D context while expanding local content. The resulting pipeline supports broader camera motion and improves consistency and visual quality in both indoor and outdoor settings. This is a credible source because it is a peer-reviewed CVPR 2024 publication with a clear technical decomposition, ablation studies, and comparative evaluation against recent text-to-3D scene baselines. The paper is relevant to this project because it provides concrete design patterns for balancing global scene coherence with iterative local generation. For a Holodeck-style system, the main takeaway is architectural: represent global structure in a stable intermediate form, then refine in stages rather than generating everything at once. That aligns with a deterministic compile pipeline where a minimally playable world appears first and fidelity upgrades arrive in later phases under explicit validation and budget constraints.