SPREAD: Spatial-Physical REasoning via geometry Aware Diffusion

Minzhang Li* Kuixiang Shao* Xuebing Li Yuyang Jiao Yinuo Bai
Hengan Zhou Sixian Shen Jiayuan Gu† Jingyi Yu†

ShanghaiTech University

{limzh2022, shaokx2025, lixb2025, jiaoyy2022, baiyn2022,
zhouha2025, shensx2024, gujy1, yujingyi}@shanghaitech.edu.cn

*Equal contribution. †Corresponding authors.

Abstract

*Automated 3D scene generation is pivotal for applications spanning virtual reality, digital content creation, and Embodied AI. While computer graphics prioritizes aesthetic layouts, vision and robotics demand scenes that mirror real-world complexity which current data-driven methods struggle to achieve due to limited unstructured training data and insufficient spatial and physical modeling. We propose **SPREAD**, a diffusion-based framework that jointly learns spatial and physical relationships through a graph transformer, explicitly conditioning on posed scene point clouds for geometric awareness. Moreover, our model integrates differentiable guidance for collision avoidance, relational constraint, and gravity, ensuring physically coherent scenes without sacrificing relational context. Our experiments on 3D-FRONT and ProcTHOR datasets demonstrate state-of-the-art performance in spatial-relational reasoning and physical metrics. Moreover, **SPREAD** outperforms baselines in scene consistency and stability during pre- and post-physics simulation, proving its capability to generate simulation-ready environments for embodied AI agents.*[1]

1. Introduction

Automated 3D scene generation is a critical task with applications ranging from virtual reality [24, 39, 51] and digital content creation [17, 28, 55] to Embodied AI [5, 23, 36]. Different domains place distinct requirements on the generated scenes. In computer graphics, the emphasis is often on geometry details, aesthetic object layouts and stylistic consistency, resulting in visually appealing and orderly environments. In contrast, 3D scenes used for training models in computer vision [9, 42, 47] and robotics [13, 57] are expected to closely mirror the complexities of the real world. These scenes must accommodate cluttered arrangements, heavy occlusions, and diverse object poses to ensure the robustness of perception systems and embodied agents. Real-world environments can often appear "chaotic". For instance, a toy played by an infant may be placed in an arbitrary pose, without any discernible logic. Capturing such diversity and disorder remains a major challenge for current data-driven methods, primarily due to the lack of sufficiently varied and unstructured training data.

Humans possess strong spatial and physical reasoning abilities that allow them to interpret and navigate the often chaotic real world. For instance, people can infer the pose of an object based on its interactions and relationships with surrounding objects. A pencil, for example, is more likely to lie flat on the ground for stability, but it can also stand upright when supported by a holder. The same object can exhibit different stable poses depending on its physical context. Beyond these physical relationships, humans also leverage deeper layers of understanding—such as functionality (e.g., a cup must remain upright to hold water) and cultural conventions (e.g., the placement of a knife and spoon to signal satisfaction with a meal). To generate truly realistic 3D scenes, a generative model must learn to reason about this foundational layer of physics: how objects stably interact, support each other, and coexist spatially in a functionally and semantically coherent manner.

Previous methods based on optimization [37, 58, 59] are capable of achieving physically plausible results for individual scenes but suffer from poor scalability. Procedural generation techniques employ handcrafted spatial rules, enabling efficient large-scale generation but introducing artificial biases that reduce real-world variability. Recent deep generative models learn scene distributions directly from data – some methods [27, 41, 52] incorporate spatial relations as graph priors from text prompts or user specifications, achieving controllability but often neglecting physi-

[1] Our code and dataset are publicly available at https://github.com/L-avenir/SPREAD.

Figure 1. **Illustration of SPREAD**, a diffusion-based framework for generating physically plausible 3D scenes with rich object interactions. (A) **SPREAD** synthesizes detailed object-level layouts with natural spatial and physical interactions, going beyond coarse layout arrangements. (B) **SPREAD** faithfully adheres to provided spatial and physical graph priors, $\mathcal{G}$. (C) **SPREAD** can provide **simulation-ready** environments for embodied AI agents.

cal constraints (leading to floating objects or penetration artifacts). Alternative approaches [53] enforce physical plausibility through guidance mechanisms, but fail to maintain realistic relational context, resulting in physically stable yet layout-incoherent scenes. Furthermore, most methods rely on datasets like 3D-FRONT [11] that capture only coarse furniture arrangements, lacking the detailed object interaction data necessary for complex physical relationships.

To this end, we introduce **SPREAD**, a guided diffusion framework that takes a foundational step toward reliable 3D scene generation by learning to reason about both spatial and physical relationships. These relationships are represented as graphs and incorporated into the diffusion process via graph transformer. Unlike prior methods [41, 53] that rely on implicit shape embeddings, our framework conditions on the posed scene point cloud at each diffusion step with a geometry-aware perceiver module. Furthermore, **SPREAD** systematically enforces fundamental physical principles—such as mesh-level collision avoidance, stable object support, and adherence to gravity—through a set of carefully designed differentiable guidance functions. This integrated design enables the generative process to satisfy key relational and physical constraints, effectively determining the placement and orientation of objects in a physically consistent manner.

Our experiments demonstrate that **SPREAD** achieves state-of-the-art performance in spatial reasoning and physical plausibility, notably exhibiting low mesh-level collision rates. The generated scenes are simulation-ready, requiring little to no post-processing to ensure physical stability. To highlight the model's capability for fine-grained spatial modeling, we train and evaluate **SPREAD** on both 3D-Front and ProcTHOR [5], the latter offering a rich diversity of small objects.

In summary, our main contributions are threefold:

- We propose a novel diffusion model that jointly represents spatial and physical relationships as differentiable graph priors, enabling scene synthesis that is both semantically coherent and physically plausible.
- We formulate mesh-level collision avoidance, stable physical relations and gravity as differentiable guidance functions, ensuring generated scenes obey physical principles while preserving rich object interactions.
- **SPREAD** achieves SOTA performance on physical plausibility metrics. Moreover, it demonstrates strong spatial and physical stability after simulations, validated on both furniture-scale and fine-grained interaction datasets.

2. Related work

2.1. 3D scene generation

Procedural Scene Generation Procedural scene generation relies on predefined rules and is widely used in indoor design [38]. Prior work used statistical methods to derive object-distribution rules for structurally sound scene gener-

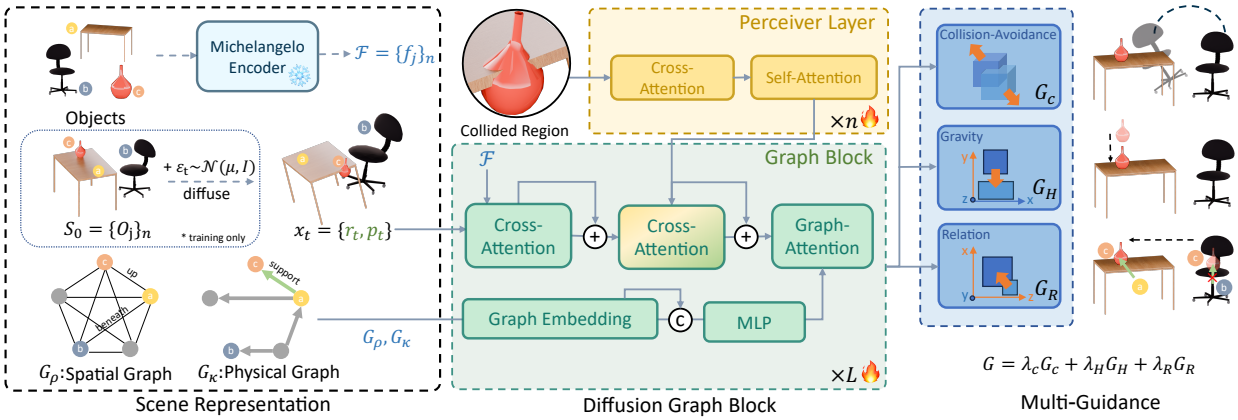

Figure 2. **Overview of SPREAD.** We propose **SPREAD**, a diffusion-based framework for generating physically plausible 3D scenes, which integrates relational constraints through spatial ($\mathcal{G}_\rho$) and physical graphs ($\mathcal{G}_\kappa$) while leveraging geometric perception via Perceiver Layers. The model employs graph-attention guided diffusion to jointly optimize physical plausibility and spatial relations during generation, producing realistic scenes with natural object interactions.

ation [2, 3, 10]. Other works cast generation as an optimization problem with constraints embedded in cost functions [5, 58]. Recently, large language models (LLMs) have introduced text-guided paradigms [8, 29, 54]. These methods rely on manual rules, limiting their use in complex scenes. Our learning-based approach models rules from data, enabling more complex scene generation.

Graph-Driven Scene Generation Graph-driven methods represent scenes as object–relationship graphs, capturing interactions and guiding construction. Semantic-spatial relation [46], dense relational [61], and hierarchical graphs [25] have been explored to capture spatial dependencies. Recent works further introduce commonsense-enhanced [56] and language-guided scene graph [27] to incorporate high-level semantics. However, existing approaches primarily focus on modeling spatial relations, often neglecting the explicit incorporation of physical relations like "support". To address this limitation, our method enhances scene priors by introducing a physical relation graph, enabling more comprehensive and physically plausible scene generation.

Diffusion-Based Scene Generation Diffusion models are a leading generative paradigm, enabling high-quality and diverse synthesis via iterative noising-denoising [16, 40]. In 3D scene generation, diffusion models enable object-level shape editing [56] and occlusion-aware inpainting [45], showing promising results. Recently, they have been applied to full-scene synthesis, modeling entire 3D scene distributions for greater flexibility and expressiveness [1, 18, 27, 41, 53]. However, existing methods tend to focus on visual quality in 3D scenes, while paying less attention to both the consistency of object relationships and physical realism, leaving the joint modeling of these two aspects relatively unexplored in diffusion-based approaches.

2.2. Guided diffusion

Guided diffusion models use external signals to steer generation, enabling fine-grained control and alignment with objectives [6, 15, 31, 35]. Recent 3D scene generation studies have adopted guided diffusion models. SceneDiffuser [21] conditions on physical constraints, while Physcene [53] integrates physical and interaction cues. Unlike Physcene, which relies on bounding-box representations, our framework leverages mesh- and relation-level guidance to enable finer-grained, physically consistent control and to produce more realistic, structurally complex scenes.

2.3. Physical constraints

Physical constraints are physics-based rules (e.g. motion laws and object interactions) that ensure plausibility in simulation. Simulation-based 3D modeling uses physics engines to enforce hard constraints for realism [30, 43, 49]. Motion capture uses physics-informed losses to enhance realism, plausibility, and temporal coherence [19, 22, 50]. In generative models, physical constraints are used as conditional signals [12] or latent embeddings [48] to enforce physical consistency. Our method guides scene generation with physical constraints, ensuring consistency and coherence for realistic, complex 3D scenes.

3. Method

To generate physically plausible scenes, we propose **SPREAD**, an integrated framework that combines our proposed geometric perceiver layers with guided diffusion. In section 3.1, we outline the compositional elements of our

scene representation, especially geometric and relational priors. In section 3.2 and section 3.3, we detail our model architecture which integrates Perceiver Layer - a dedicated geometric perception module that enables the network to learn geometric constraints during training. In section 3.4, for posterior optimization during inference, we propose a novel combination of diffusion guidance mechanism that simultaneously addresses physical plausibility and relational constraints.

3.1. Scene representation

To enable comprehensive modeling and generation of physically plausible scenes, our framework relies on structured representations that precisely capture objects and their spatial interactions as priors.

As illustrated in Fig 2, the scene S_i contains objects o_j^i each defined by a tuple $\langle p_j^i, r_j^i, f_j^i, \rho_j^i, \kappa_j^i \rangle$, where 3D translation $p_j^i \in \mathbb{R}^3$ stands for the centroid position of o_j^i, orientation $r_j^i \in \mathtt{SO}(3)$. Moreover, the geometric features $f_j^i \in \mathbb{R}^d$ provide a d-dimensional shape descriptor. The spatial relations ρ_j^i model pairwise relative directions. For example, $\rho(o_k^i, o_l^i)$ indicates if o_l^i is left, right, front, or back of o_k^i. Similarly, the physical interactions $\kappa_j^i(o_k^i, o_l^i)$ describe support, contact, or attachment relation between objects, which enables the generation with more comprehensive modeling and additional controllability. Notably, instead of using images as visual references [4, 26], we employ explicitly structured graph representations to separately model spatial relationships $\mathcal{G}_\rho \in \{0, ..., m\}^{N \times N}$ and physical interactions $\mathcal{G}_\kappa \in \{0, 1, ..., q\}^{N \times N}$ as latent constraints, where m and q represent the number of spatial relations and physical interactions respectively.Technically, we utilize a continuous representation [14] to parameterize rotation r, where $r \in \mathbb{R}^6$.

3.2. Geometry-aware diffusion modeling

Here, we introduce **SPREAD**, our graph-based diffusion for scene generation with geometric awareness. As illustrated in Fig 2, **SPREAD** differs from existing scene generation methods by explicitly modeling physical and spatial relations between denoised meshes at each denoising step, while incorporating additional geometric inputs through a geometry-aware perceiver module. Furthermore, we employ a graph transformer with cross-attention blocks to parameterize $\epsilon_\theta(\mathbf{x}_t, t, \mathbf{f}, \mathcal{G}_\rho, \mathcal{G}_\kappa)$, where t denotes time embeddings, $\mathbf{f}$ denotes geometric features for collections of objects in the scene, $\mathcal{G}_\rho$ and $\mathcal{G}_\kappa$ represent spatial and interaction relation graph, respectively.

The diffusion process operates on a structured scene representation $\mathcal{S}_i = \{o_j^i\}_{j=1}^N$, where each object is parameterized as $\langle p_j^i, r_j^i, f_j^i, \rho_j^i, \kappa_j^i \rangle$. We construct a joint state space by concatenating scene representations across all objects,

$$\mathbf{x}_0 = \bigoplus_{j=1}^{N} [p_j^i \| r_j^i] \in \mathbb{R}^{N \times (3+6)} \tag{1}$$

forming the basis for the diffusion process. The forward process follows a Markov chain that gradually perturbs the data through Gaussian transitions, preserving the topological structure while adding noise

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}) \tag{2}$$

where β_t stands for the variance of the Gaussian noise added at each step of the forward (diffusion) process.

For reversal process, we model the spatial relations $\mathcal{G}\rho$ and physical relations $\mathcal{G}\kappa$ between objects as graph structures, where each graph is represented by an adjacency matrix of shape (N, N), with elements indicating K possible relation types. These graph structures are first mapped to a continuous latent space through embedding layers,

$$\mathbf{E} = \text{MLP}(\text{Embedding}(\mathcal{G})) \tag{3}$$

where edge embeddings $\mathbf{E} \in \mathbb{R}^{N \times N \times d_e}$. These embeddings are then injected as bias terms into the graph attention layers. At each denoising step t, the graph structures condition the generation process through the diffusion graph block

$$\mathbf{H}t^{l+1} = \text{GraphBlock}^l(\mathbf{H}t^l, \mathcal{G}\rho, \mathcal{G}\kappa) \tag{4}$$

while simultaneously processing both types of relational information.

By enabling relational modeling in continuous feature space through MLP projection, **SPREAD** achieves joint optimization of spatial relations and physical constraints in each graph block layer.

3.3. Model architecture

Our geometry-aware diffusion network jointly models object orientation, geometry, and inter-object relations as multimodal priors to transform a random layout into a physically plausible, semantically coherent 3D scene. As described in section 3.2, each object node is represented by a 9-dimensional state vector $\mathbf{x}_i$. This vector is then augmented with fixed sinusoidal positional encodings and projected into the attention dimension through a linear layer. Concurrently, each object is encoded by a pretrained Michelangelo encoder [60] into 256 64-dimensional tokens, which are then linearly projected to the attention dimension. These shape tokens remain constant during diffusion, providing a stable shape prior.

At each diffusion timestep t, dynamic geometric interactions are captured by sampling M points $\mathbf{p}_i^M$ on the noisy mesh of object i and computing the one-way Chamfer distance [7] to all other objects' point clouds $\mathcal{P}_{\neg i}$. We assign a

Table 1. **Quantitative comparison.** Our method matches baseline FID on 3D-FRONT while setting new state-of-the-art on ProcTHOR: it dramatically reduces mesh collisions, achieves the highest graph recall (GRecall), minimizes average support distance (ASD), and delivers the greatest scene stability under Isaac Sim.

Method	3D-FRONT				ProcTHOR				
	$\text{Col}_{\text{mesh}} \downarrow$			FID↓	GRecall↑	$\text{Col}_{\text{mesh}} \downarrow$	ASD↓	Stability↑	FID↓
	Bedroom	Livingroom	Diningroom						
ATISS	0.275	0.451	0.428	68.0	/	0.174	0.510	0.813	33.9
DiffuScene	0.298	0.359	0.376	61.6	/	0.360	0.071	0.886	21.4
InstructScene	0.285	0.350	0.331	**61.3**	0.964	0.260	0.021	0.876	20.0
Ours	**0.097**	**0.185**	**0.183**	64.7	**0.979**	**0.121**	**0.007**	**0.950**	**18.8**

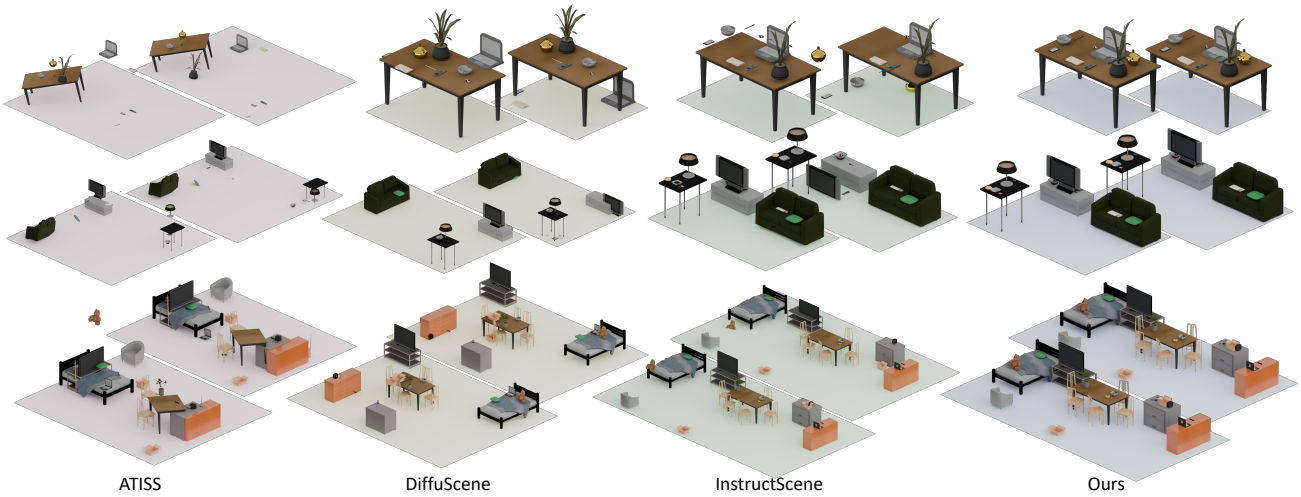

Figure 3. **Comparative Generation and Simulation Results.** Visual comparison of scene layouts produced by our method versus three baseline approaches, shown before (left) and after (right) physics simulation.

sign to each distance via the nearest neighbor's normal $\mathbf{n}_{\text{nn}}$, thereby approximating a signed distance field:

$$d_{\text{scd}}(\mathbf{p}) = \min_{\mathbf{q} \in \mathcal{P}_{\neg i}} \|\mathbf{p} - \mathbf{q}\|_2 \cdot \text{sign}\big(\mathbf{n}_{\text{nn}}^{\top}(\mathbf{p} - \mathbf{q})\big) \tag{5}$$

This procedure yields a feature tensor of shape $(B, N, M, 4)$, where the first three channels encode global coordinates and the fourth channel encodes d_{scd}. A Perceiver [20] module then distills these sparse, high-dimensional features into n d-dimensional tokens $\mathbf{f}^{\text{geo}}$ via cross-attention, enabling the network to perceive collisions and penetrations.

Discrete spatial and physical relations (e.g., "left of," "supports") are embedded and concatenated as edge features $\mathbf{e}_{\rho\kappa}$, together with node features forming an explicit scene graph. We stack L multimodal graph layers to iteratively fuse and propagate information: within each block, node features first attend to static shape tokens and then to dynamic geometric embeddings via sequential cross-attention, producing shape- and geometry-aware fused representations; these representations and the edge features are subsequently processed by a graph attention block, in which multi-head graph attention propagates signals along explicit edges. All normalization layers condition on the timestep embedding $\mathbf{t}_{\text{emb}}$ via AdaLayerNorm [33, 34], making the denoising process time-aware. Finally, the refined node features are projected by a MLP to predict the noise $\hat{\epsilon}$, and training minimizes the mean squared error $\|\hat{\epsilon} - \epsilon\|^2$. By deeply integrating static shape priors, dynamic Chamfer-based geometry, and explicit relational structure, our framework achieves high-fidelity, physically consistent 3D scene generation.

3.4. Multi-guidance framework

The proposed diffusion guidance framework incorporates physically-grounded constraints through differentiable operators that modify the score function of the diffusion process. Formally, given a diffusion model with learned score function $s_\theta(\mathbf{x}_t, t)$, the guided reverse process follows:

$$\nabla_{\mathbf{x_t}} \log p_\gamma(\mathbf{x_t}) = s_\theta(\mathbf{x_t}, t) + \gamma \nabla_{\mathbf{x_t}} \mathcal{G}(\mathbf{x_t}) \tag{6}$$

where $\mathcal{G}(\mathbf{x_t})$ represents our composite guidance signal combining three key components: collision guidance, grav-

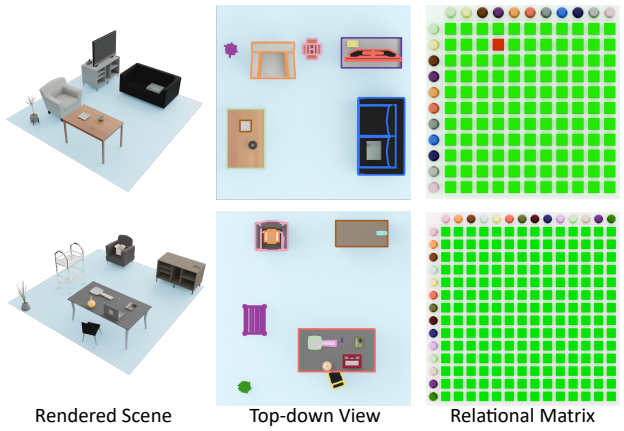

Figure 4. **Scene & Relation Visualization.** For two generated scenes, we show the final render (left), the top-down layout (middle), and the pairwise relation evaluation matrix (right). The matrix encodes every object-pair's spatial relation: green entries denote correct relations (w.r.t. the ground truth), and red entries denote incorrect ones.

ity guidance, and relation guidance. With the weight of each guidance, the overall guidance is defined via:

$$\mathcal{G} = \lambda_C \mathcal{G}_C + \lambda_H \mathcal{G}_H + \lambda_R \mathcal{G}_R \tag{7}$$

Together, these guidance terms enforce consistency with real-world physical laws and foster the generation of stable, contact-rich 3D scenes.

Collision Guidance We introduce a collision-avoidance objective function based on intersecting triangles from different meshes. Unlike approaches like Physcene [53] which take predicted bounding boxes as approximation, our method directly quantifies collision relationships based on mesh triangles, which proves more efficient and accurate. Specifically, we use a and b to represent any two distinguish objects in the scene. And we use t_a^i and t_b^j to denote the i-th triangle of a and the j-th triangle of b. We use **CoDF** [44] to evaluate the mesh-based collision-free guidance via:

$$\mathcal{G}_C = \frac{1}{|C|} \sum_{a,b,a\neq b} \sum_{(i,j)\in C} \mathbf{CoDF}(t_a^i, t_b^j) \tag{8}$$

where C represents collision triangle pairs found with BVH, and **CoDF** represents the conical distance field. The overall penalty is obtained by summing the penalties with all collision pairs of surface patches from different objects.

Gravity Guidance To ensure physically plausible support relationships, we model gravity constraints by computing vertical distances between objects and their supporters. For each object, we compute the vertical offset $r_i = d_i - \epsilon$,

Table 2. **Ablation on ProcTHOR.** starting from our base model, adding the geometry module and then the full multi-guidance framework yields consistent improvements in all physical metrics.

Method	GRecall↑	Col_{mesh} ↓	ASD↓	Stability↑
Ours	0.963	0.241	0.014	0.934
+Geometry	0.965	0.225	0.012	0.938
+Guidance	**0.979**	**0.121**	**0.007**	**0.950**

where d_i is the distance from the object to its supporter, and ϵ is an empirical minimal threshold preventing objects from potential intersection caused by gravity guidance. This formulation naturally penalizes both excessive floating ($r_i > \theta_H$) and interpenetration ($r_i < 0$) while allowing small deviations within tolerance θ_H.

$$\mathcal{G}_H = \sum_{r_i > \theta_H \ \vee \ r_i < 0} |r_i| \tag{9}$$

Relation Guidance In reality, objects exhibit complex interrelations. To model these, we introduce a score function based on the extent of overlap between their projections onto the XZ-plane. This approach is both effective and efficient, since we approximate each object by its projection convex hull. Instead of calculating the exact overlap area, we estimate a penalty by measuring the distances from all the object vertices lying outside their supporting object to the supporting hull. To be more specific, we denote a directed pair (i, j) to indicate that the object i is supported by the object j, and we use E to represent the set of all such relations. Additionally, we denote $V_{i,j}$ to represent all the vertices of the i-th object's projection outside the convex hull of the j-th object. Thus, we further define the guidance as follows:

$$\mathcal{G}_R = \sum_{(i,j)\in E} \sum_{\alpha \in V_{i,j}} \frac{\mathbf{s}(\alpha, j)}{|V_{i,j}||E|} \tag{10}$$

where $\mathbf{s}(\alpha, j)$ represents the minimal Euclidean distance between a point α to the convex hull of supporter j.

4. Experiments

Datasets We evaluate our model on two large-scale indoor datasets: (1) 3D-FRONT [11] and (2) ProcTHOR [5], which together capture both aesthetic layouts and rich physical interactions. 3D-FRONT provides high-quality, designer-curated scenes; we adopt the InstructScene [27] preprocessing pipeline, augmenting each scene with explicit relative-position annotations (e.g., "left-of", "above") to form structured scene graphs. To better model complex object interactions, we utilize ProcTHOR's procedurally generated indoor scenes, excluding non-supportive meshes

(e.g., wall hangings). We then apply physics-based corrections to resolve interpenetration and remove floating objects, while annotating spatial relationships using the same scheme as in 3D-FRONT.

Baselines We compare **SPREAD** against three baselines: (1) ATISS [32], a permutation-invariant transformer that models scene generation as an autoregressive process over an unordered set of objects. (2) DiffuScene [41], a model that adopts a denoising diffusion probabilistic model to generate scenes in a non-autoregressive manner. (3) InstructScene, which introduces a two-stage, graph-based framework designed for instruction-driven synthesis. These baselines provide a robust benchmark covering autoregressive, diffusion-based, and graph-structured generative methodologies. We refer the reader to the Supplementary Material for more information.

Metrics To ensure a comprehensive evaluation of our model, we assess visual quality, physical plausibility, and structural fidelity. For visual fidelity and diversity, we use the *Fréchet Inception Distance* (FID). Physical plausibility is measured via the *Mesh Collision Rate*, which quantifies object intersections. For the ProcTHOR dataset, we introduce three specialized metrics: *Graph Recall* (GRecall) to evaluate structural accuracy by comparing inferred spatial relationships, *Average Support Distance* (ASD) to assess contact surface quality through signed distance functions, and *Stability*, which is measured by simulating scenes in NVIDIA Isaac Sim [30] and verifying object relationship consistency. Metric details are attached in the supplementary material.

4.1. Results on scene generation

Our framework's advantages become especially clear when evaluated on the interaction-rich ProcTHOR dataset. As Table 1 shows, although we match baseline FID on 3D-FRONT, we surpass all existing approaches on every physically grounded metric in ProcTHOR. In particular, our method dramatically reduces Col_{mesh} and achieves a GRecall of 0.979—demonstrating faithful adherence to the true scene layout (Figure 4). It yields an Average Support Distance of just 0.007, indicating virtually gap-free contact.

4.2. Ablation study

Our ablation study on the ProcTHOR dataset evaluates the separate contributions of the geometry-aware perceiver module and multi-guidance framework (Table 2). The baseline vanilla diffusion model shows substantial improvement when augmented with the geometry-aware perceiver (+Geometry), evidenced by reduced collision rate and ASD metrics. The complete multi-guidance framework (+Guidance) combined with geometry-aware diffusion achieves the most significant gains, reaching state-of-the-art physical plausibility. As Figure 5 illustrates: collision guidance effectively removes interpenetrations, gravity guidance eliminates floating artifacts, and relational guidance maintains proper support structures - collectively achieving the lowest ASD, highest GRecall, and best simulation stability.

Table 3. **Comparison of inference times** (in seconds) across different scene generation methods.

Method	**SPREAD**	ATISS	InstructScene	DiffuScene
Inference Time (s)	14.72	0.02	2.58	10.25

4.3. Scene consistency in pre-post simulation

To rigorously evaluate physical plausibility, we measure scene stability under NVIDIA Isaac Sim. As Table 1 shows, our method achieves an Isaac Stability score of 0.950, meaning the vast majority of pairwise object relationships remain intact after simulation. Figure 3 offers a side-by-side comparison: while baseline layouts frequently suffer object displacement and structural drift under physical forces, ours remain virtually unchanged, underscoring the practical benefit of integrating physical reasoning directly into the generation process.

4.4. User study

We have conducted a user study focusing on physical consistency, assessing adherence to physical laws, and scene rationality, evaluating high-level semantic and commonsense coherence. We randomly selected five scenes, each containing a minimum of 10 and a maximum of 22 objects. A total of 57 valid responses were collected, each evaluating randomly selected three scenes to determine the preferred method. As illustrated in Figure 6, our method garnered 88.6% of votes, thus underscoring its superiority over ATISS(0.9%), DiffuScene(6.1%) and InstructScene(4.4%). This finding suggests that our method generates scenes that are more physically consistent, scene-rational, and show greater consistency with human preferences.

4.5. Inference speed comparison

In this section, we analyze the computational efficiency of state-of-the-art methods. The average inference latency per scene is as follows: ATISS (0.02s), InstructScene (2.58s), Diffuscene (10.25s), and SPREAD (14.72s). The higher latency of our method is an inherent trade-off of its generative architecture, deliberately designed to model complex object relationships and generate coherent scenes, prioritizing quality and structural integrity over inference speed.

Figure 5. **Guidance Ablation.** Results showing effect of different guidance terms. Each major row compares results before (top) and after (bottom) adding a specific guidance. Columns show different scenes. Red circles highlight issues such as collisions, floating, or incorrect spatial relations before guidance; green circles show improvements after applying guidance, with zoom-in views for clarity.

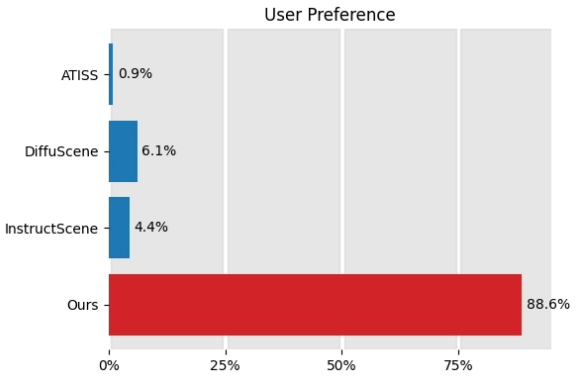

Figure 6. **User Study.** Our method dominated with 88.6% of the votes, above ATISS with 0.9%, InstructScene with 4.4%, and DiffuScene with 6.1%, indicating that our method better preserves physical consistency and scene rationality.

5. Conclusion

We present **SPREAD**, a guided diffusion framework that jointly models object spatial and physical relationships through differentiable graph priors and multi-guidance mechanism. Our method synthesizes physically plausible and simulation-ready scenes. Experiments demonstrate superior performance in spatial reasoning and physical metrics, with robustness validated under simulations.

Future work While effective at preserving physical constraints, our method is currently limited to indoor scenes due to dataset availability. Future work will extend to outdoor generation by leveraging image-conditioned paradigms [29, 55]. To address the computational cost inherent in our iterative diffusion process, we will explore efficient alternatives like flow matching. Furthermore, we will investigate formulating the diffusion directly on the SE(3) manifold to better leverage its geometric prior for more principled scene synthesis.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant W2431046, National Key R&D Program of China 2025YFA1309603, Central Guided Local Science and Technology Foundation of China YDZX20253100001001, and by MoE Key Lab of Intelligent Perceptionand Human-Machine Collaboration (ShanghaiTech University), the Shanghai Frontiers Science Center of Human-centered Artificial Intelligence. The experiments of this work were supported by the SIST computing Platform and HPC, ShanghaiTech University.

References

[1] Miguel Angel Bautista, Pengsheng Guo, Samira Abnar, Walter Talbott, Alexander Toshev, Zhuoyuan Chen, Laurent Dinh, Shuangfei Zhai, Hanlin Goh, Daniel Ulbricht, et al. Gaudi: A neural architect for immersive 3d scene generation. *Advances in Neural Information Processing Systems*, 35:25102–25116, 2022. 3

[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

[3] Angel X Chang, Mihail Eric, Manolis Savva, and Christopher D Manning. Sceneseer: 3d scene design with natural language. *arXiv preprint arXiv:1703.00050*, 2017. 3

[4] Jit Chatterjee and Maria Torres Vega. 3d-scene-former: 3d scene generation from a single rgb image using transformers. *The Visual Computer*, 41(4):2875–2889, 2025. 4

[5] 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. *Advances in Neural Information Processing Systems*, 35:5982–5994, 2022. 1, 2, 3, 6

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 3

[7] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017. 4

[8] 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. 3

[9] Michael Fischer, Iliyan Georgiev, Thibault Groueix, Vladimir G Kim, Tobias Ritschel, and Valentin Deschaintre. Sama: Material-aware 3d selection and segmentation. *arXiv preprint arXiv:2411.19322*, 2024. 1

[10] Matthew Fisher and Pat Hanrahan. Context-based search for 3d models. In *ACM SIGGRAPH Asia 2010 papers*, pages 1–10. 2010. 3

[11] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10933–10942, 2021. 2, 6, 1

[12] Nate Gillman, Charles Herrmann, Michael Freeman, Daksh Aggarwal, Evan Luo, Deqing Sun, and Chen Sun. Force prompting: Video generation models can learn and generalize physics-based control signals. *arXiv preprint arXiv:2505.19386*, 2025. 3

[13] Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, et al. Maniskill2: A unified benchmark for generalizable manipulation skills. *arXiv preprint arXiv:2302.04659*, 2023. 1

[14] Thorsten Hempel, Ahmed A Abdelrahman, and Ayoub Al-Hamadi. 6d rotation representation for unconstrained head pose estimation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2496–2500. IEEE, 2022. 4

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[17] Lukas Höllein, 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*, pages 7909–7920, 2023. 1

[18] Siyi Hu, Diego Martín Arroyo, Stephanie Debats, Fabian Manhardt, Luca Carlone, and Federico Tombari. Mixed diffusion for 3d indoor scene synthesis. *ArXiv*, abs/2405.21066, 2024. 3

[19] Buzhen Huang, Liang Pan, Yuan Yang, Jingyi Ju, and Yangang Wang. Neural mocon: Neural motion control for physically plausible human motion capture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6417–6426, 2022. 3

[20] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 5

[21] Max Jiang, Yijing Bai, Andre Cornman, Christopher Davis, Xiukun Huang, Hong Jeon, Sakshum Kulshrestha, John Lambert, Shuangyu Li, Xuanyu Zhou, et al. Scenediffuser: Efficient and controllable driving simulation initialization and rollout. *Advances in Neural Information Processing Systems*, 37:55729–55760, 2024. 3

[22] Jingyi Ju, Buzhen Huang, Chen Zhu, Zhihao Li, and Yangang Wang. Physics-guided human motion capture with pose probability modeling. *arXiv preprint arXiv:2308.09910*, 2023. 3

[23] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Wensi Ai, Benjamin Martinez, et al.

Behavior-1k: A human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation. *arXiv preprint arXiv:2403.09227*, 2024. 1

[24] Ke Li, Tim Rolff, Susanne Schmidt, Reinhard Bacher, Simone Frintrop, Wim Leemans, and Frank Steinicke. Immersive neural graphics primitives. *arXiv preprint arXiv:2211.13494*, 2022. 1

[25] Manyi Li, Akshay Gadi Patil, Kai Xu, Siddhartha Chaudhuri, Owais Khan, Ariel Shamir, Changhe Tu, Baoquan Chen, Daniel Cohen-Or, and Hao Zhang. Grains: Generative recursive autoencoders for indoor scenes. *ACM Transactions on Graphics (TOG)*, 38(2):1–16, 2019. 3

[26] Ming-Feng Li, Yueh-Feng Ku, Hong-Xuan Yen, Chi Liu, Yu-Lun Liu, Albert YC Chen, Cheng-Hao Kuo, and Min Sun. Genrc: Generative 3d room completion from sparse image collections. In *European Conference on Computer Vision*, pages 146–163. Springer, 2024. 4

[27] Chenguo Lin and Yadong Mu. Instructscene: Instruction-driven 3d indoor scene synthesis with semantic graph prior. *arXiv preprint arXiv:2402.04717*, 2024. 1, 3, 6

[28] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 300–309, 2023. 1

[29] 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. 3, 8

[30] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021. 3, 7

[31] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3

[32] Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. Atiss: Autoregressive transformers for indoor scene synthesis. *Advances in Neural Information Processing Systems*, 34:12013–12026, 2021. 7, 3

[33] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 5

[34] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 5

[35] Nicholas Ezra Pfaff, Hongkai Dai, Sergey Zakharov, Shun Iwase, and Russ Tedrake. Steerable scene generation with post training and inference-time search. In *9th Annual Conference on Robot Learning*, 2025. 3

[36] Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, et al. Habitat 3.0: A co-habitat for humans, avatars and robots. *arXiv preprint arXiv:2310.13724*, 2023. 1

[37] Siyuan Qi, Yixin Zhu, Siyuan Huang, Chenfanfu Jiang, and Song-Chun Zhu. Human-centric indoor scene synthesis using stochastic grammar. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5899–5908, 2018. 1

[38] Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, et al. Infinigen indoors: Photorealistic indoor scenes using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21783–21794, 2024. 2

[39] Anton Ratnarajah and Dinesh Manocha. Listen2scene: Interactive material-aware binaural sound propagation for reconstructed 3d scenes. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pages 254–264. IEEE, 2024. 1

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[41] Jiapeng Tang, Yinyu Nie, Lev Markhasin, Angela Dai, Justus Thies, and Matthias Nießner. Diffuscene: Denoising diffusion models for generative indoor scene synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20507–20518, 2024. 1, 2, 3, 7

[42] Anh Thai, Weiyao Wang, Hao Tang, Stefan Stojanov, James M Rehg, and Matt Feiszli. 3× 2: 3d object part segmentation by 2d semantic correspondences. In *European Conference on Computer Vision*, pages 149–166. Springer, 2024. 1

[43] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012. 3

[44] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision*, 118(2): 172–193, 2016. 6

[45] Guangcong Wang, Peng Wang, Zhaoxi Chen, Wenping Wang, Chen Change Loy, and Ziwei Liu. Perf: Panoramic neural radiance field from a single panorama. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10): 6905–6918, 2024. 3

[46] Kai Wang, Yu-An Lin, Ben Weissmann, Manolis Savva, Angel X Chang, and Daniel Ritchie. Planit: Planning and instantiating indoor scenes with relation graph and spatial prior networks. *ACM Transactions on Graphics (TOG)*, 38(4):1–15, 2019. 3

[47] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of

novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17868–17879, 2024. 1

[48] Qianyang Wu, Ye Shi, Xiaoshui Huang, Jingyi Yu, Lan Xu, and Jingya Wang. Thor: Text to human-object interaction diffusion via relation intervention. *arXiv preprint arXiv:2403.11208*, 2024. 3

[49] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11097–11107, 2020. 3

[50] Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. Physics-based human motion estimation and synthesis from videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11532–11541, 2021. 3

[51] Linning Xu, Vasu Agrawal, William Laney, Tony Garcia, Aayush Bansal, Changil Kim, Samuel Rota Bulò, Lorenzo Porzi, Peter Kontschieder, Aljaž Božič, et al. Vr-nerf: High-fidelity virtualized walkable spaces. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–12, 2023. 1

[52] Xiuyu Yang, Yunze Man, Junkun Chen, and Yu-Xiong Wang. Scenecraft: Layout-guided 3d scene generation. *Advances in Neural Information Processing Systems*, 37: 82060–82084, 2024. 1

[53] Yandan Yang, Baoxiong Jia, Peiyuan Zhi, and Siyuan Huang. Physcene: Physically interactable 3d scene synthesis for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16262–16272, 2024. 2, 3, 6

[54] Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, et al. Holodeck: Language guided generation of 3d embodied ai environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16227–16237, 2024. 3

[55] Kaixin Yao, Longwen Zhang, Xinhao Yan, Yan Zeng, Qixuan Zhang, Lan Xu, Wei Yang, Jiayuan Gu, and Jingyi Yu. Cast: Component-aligned 3d scene reconstruction from an rgb image. *ACM Transactions on Graphics (TOG)*, 44(4): 1–19, 2025. 1, 8

[56] Guangyao Zhai, Evin Pınar Örnek, Shun-Cheng Wu, Yan Di, Federico Tombari, Nassir Navab, and Benjamin Busam. Commonscenes: Generating commonsense 3d indoor scenes with scene graph diffusion. *Advances in Neural Information Processing Systems*, 36:30026–30038, 2023. 3

[57] Jialiang Zhang, Haoran Liu, Danshi Li, XinQiang Yu, Haoran Geng, Yufei Ding, Jiayi Chen, and He Wang. Dexgraspnet 2.0: Learning generative dexterous grasping in large-scale synthetic cluttered scenes. In *8th Annual Conference on Robot Learning*, 2024. 1

[58] Shao-Kui Zhang, Yi-Xiao Li, Yu He, Yong-Liang Yang, and Song-Hai Zhang. Mageadd: Real-time interaction simulation for scene synthesis. In *Proceedings of the 29th ACM international conference on multimedia*, pages 965–973, 2021. 1, 3

[59] Yizhou Zhao, Kaixiang Lin, Zhiwei Jia, Qiaozi Gao, Govind Thattai, Jesse Thomason, and Gaurav S Sukhatme. Luminous: Indoor scene generation for embodied ai challenges. *arXiv preprint arXiv:2111.05527*, 2021. 1

[60] Zibo Zhao, Wen Liu, Xin Chen, Xianfang Zeng, Rui Wang, Pei Cheng, Bin Fu, Tao Chen, Gang Yu, and Shenghua Gao. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. *Advances in neural information processing systems*, 36:73969–73982, 2023. 4, 1

[61] Yang Zhou, Zachary While, and Evangelos Kalogerakis. Scenegraphnet: Neural message passing for 3d indoor scene augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7384–7392, 2019. 3