

# SPREAD: Spatial-Physical REasoning via geometry Aware Diffusion

## Supplementary Material

### 6. Implementation details

#### 6.1. Hyper parameters

All experiments were conducted on NVIDIA H20 GPUs. The model was trained for 2,000 epochs with a batch size of 16, taking approximately 50 GPU hours. We employ the AdamW optimizer with a learning rate of  $1 \times 10^{-4}$  and a weight decay of  $1 \times 10^{-4}$ . Our diffusion backbone adopts the Denoising Diffusion Probabilistic Models (DDPM) framework. We utilize `squaredcos_cap_v2` noise schedule, operating over a total of 1,000 denoising timesteps. As mentioned in Section 3, we sample  $M$  points from mesh faces with noisy rotation and translation to compute the signed chamfer distance, which yields feature vectors during training. This strategy significantly enhances the model’s capacity to capture geometric details and contributes to spatial-physically reasoning and scene generation. Here we empirically set  $M$  to be 2,000. During sampling, we integrate three distinct guidance terms to steer the generation towards physically plausible and relationally consistent scenes. We empirically set our guidance weights  $\lambda_C, \lambda_R, \lambda_H$  for collision, relation, gravity guidance to be  $7.5 \times 10^{-3}, 1.0 \times 10^{-3}, 1.0 \times 10^{-3}$ , respectively.

During the quantitative evaluation, we compute the Fréchet Inception Distance (FID) on CLIP features using the Clean-FID<sup>2</sup> library. We employ Nvidia Isaac Sim with version 5.1.0 to evaluate simulation stability.

#### 6.2. Shape Encoder

We leverage the pre-trained shape encoder, **Michelangelo** [60], to extract latent shape codes for CAD models from 3D-FRONT [11] dataset and the processed ProcTHOR [5] dataset. It functions as an image-text-aligned 3D shape Variational Auto-Encoder. Here we briefly introduce the forward process of the encoder. First, we initially sample a point cloud  $P \in \mathbb{R}^{N \times 3}$  from the surface of the 3D shape. Then the points are fed into the **SITA-VAE** module of **Michelangelo**, which captures both low-level geometric information via  $L$  query tokens and high-level semantic information via a single global head token. Finally, multiple self-attention layers iteratively refine the representation and obtain the final embeddings  $f \in \mathbb{R}^{N \times L \times D}$ . The embeddings are used directly for the following diffusion training, serving as a compressed and semantically rich geometry representation of shapes.

#### 6.3. Geometry-aware module in diffusion

In the main paper, we introduced a novel training strategy that endows diffusion models with explicit awareness of geometry and inter-relationship between paired objects in the scene. During the diffusion training process, we aim at enabling the model to perceive collisions and penetrations in the scene. To achieve this, we sample  $M=2,000$  points from the surface for each noisy-posed object and compute the one-way Chamfer distance to all other objects’ point clouds. This procedure divides the points into two subsets: the collided points and the remains. Rather than relying on a binary collision indicator  $\{0, 1\}$ , we preserve the signed distance value as a continuous feature. Consequently, a feature tensor of shape  $(N, M, 4)$  is obtained, which represents the collision information (referred to as relational information) for the scene at timestep  $t$ . Then a lightweight yet effective perceiver-based transformer is employed to convey the explicit geometric information of the scene at timestep  $t$ .

### 7. Dataset

#### 7.1. The insufficiency of public dataset

While existing 3D scene datasets such as 3D-FRONT and ProcTHOR have made significant strides in scale and quality, they share a fundamental limitation: a pronounced lack of rich, fine-grained object-to-object relations. The 3D-FRONT dataset primarily consists of designer-curated indoor scenes, where layouts are often aesthetically oriented and orderly. This results in oversimplified spatial relationships (e.g., "against a wall" or "centered in a room") and fails to capture the complex, unstructured interactions commonly found in real-world environments, such as object stacking, partial occlusion, or casual support. Although the ProcTHOR dataset offers greater environmental diversity through procedural generation and includes more small-scale objects, its automated process does not explicitly or comprehensively annotate crucial physical support relations or precise spatial relations. This inherent sparsity of relational data makes it difficult for models trained directly on these datasets to learn and reason about the complex spatial and physical interactions characteristic of real-world scenes. Consequently, models frequently exhibit fine-grained physical inconsistencies, such as floating objects, inter-penetrations, and implausible support structures. This critical shortcoming motivates our work to not only rely on raw data but also to introduce a comprehensive pre-processing pipeline (detailed in Section 7.2) that explicitly derives these missing relations, thereby compensating for

<sup>2</sup><https://pypi.org/project/clean-fid/>

the inherent limitations of the source data and providing a foundation for learning robust spatial-physical reasoning. Here we present the comparison of our preprocessed dataset and the original 3D-FRONT dataset in Fig. 7.

## 7.2. Preprocessing

We conduct experiments on the **ProcTHOR-10K** dataset, a large-scale collection of 10,000 procedurally generated 3D indoor environments developed by the Allen Institute for AI. We implement a comprehensive, parallelized preprocessing pipeline using Blender for geometric operations.

The pipeline executes the following automated steps for each scene. First, the full house layout is deconstructed into individual rooms based on architectural boundaries. Within each room, non-supporting objects such as wall art and small decorative items are filtered out, while functional furniture such as tables, shelves, and countertops are retained. To ensure physical plausibility, the pipeline automatically rectifies geometric issues, including resolving mesh intersections between objects and grounding floating objects onto their nearest underlying supporting surfaces. Subsequently, supported objects are identified and paired with their corresponding supporting objects (e.g., a Mug on a Table). Relative spatial relationships (e.g., left, right) are computed and annotated for all object pairs.

Finally, the structured data, including object poses, meshes, and their annotated relationships, is serialized into .npz format for efficient loading during training.

Table 4. **Dataset Statistics.** Summary of the processed ProcTHOR-10K dataset, detailing the scale of rooms, splits, and object diversity.

Metric	Value
Total Processed Rooms	23,472
Training Set Size (rooms)	22,472
Test Set Size (rooms)	1,000
Unique Object Categories	108
Objects per Room (range)	2 - 52

## 8. Explanation for metrics

To provide a multifaceted evaluation of our method, we employ a suite of metrics assessing visual fidelity, physical plausibility, and relational accuracy.

**Fréchet Inception Distance (FID↓).** We assess the visual quality and diversity of our generated scenes by rendering top-down 2D views and calculating the FID score against rendered images from the ground-truth test set. This measures the distributional similarity between generated and real scene layouts. However, this metric is limited in its applicability to our method, as our approach focuses more on physical plausibility, which the FID score cannot reflect.

**Mesh Collision Rate (Col<sub>mesh</sub>↓).** This metric quantifies the degree of inter-penetration between objects. For each scene, we use `trimesh.CollisionManager` to identify all intersecting mesh pairs. We consider a collision significant if the penetration depth exceeds a threshold of 0.01 units, ignoring minor surface contacts. The final score reports the percentage of objects involved in at least one significant collision across the entire test set. A lower score indicates superior physical plausibility.

**Graph Recall (GRecall↑).** To measure structural accuracy, we evaluate the adherence of generated scenes to the input spatial-relation graph. For each generated scene, we derive the pairwise relative spatial relationships between all objects based on their final poses. GRecall measures the proportion of relationships in the ground-truth input graph that are correctly realized in the generated scene.

**Average Support Distance (ASD↓).** This metric evaluates the physical quality of support relationships. For each supported object, we compute the signed distances from its vertices to the supporting object. The absolute value of the minimum signed distance—representing the minimum gap or largest penetration—is taken as the support distance. ASD is the average of these distances over all support pairs. A lower value signifies more precise, high-quality contact.

**Isaac Stability (Stability↑).** To rigorously test physical robustness, we perform ten 100-step physics simulations for each generated scene in NVIDIA Isaac Sim. We record the poses of all objects before and after the simulation. Stability is measured as the percentage of pairwise spatial relationships that remain unchanged after the simulation concludes. A high stability score indicates that the generated layout is physically sound and resilient to gravity-induced collapse.

## 9. Ablation studies

In the main paper, we discussed the effectiveness of each design in **SPREAD**, covering model architecture and guidance choice. Here, we present further methodological details and ablation analysis.

### What is the effect of choosing Michelangelo encoder?

While conventional encoders, such as those based on open-shape or trained on small datasets, can characterize shape, they often only offer simple image-text alignment semantic features or object identifiers specific to their small dataset. When selecting a shape encoder to extract geometric features, we consider not only the compatibility between the shape embedding and the relational input embeddings but also its capacity for capturing complex underlying geometric information. The heightened requirements prompt us to employ a special encoder capable of both aligning the shape latent space with semantic modalities and capturing underlying geometric information. Our early experiments revealed modest improvement compared with the plain shape encoder in performance metrics like the rate of mesh colli-

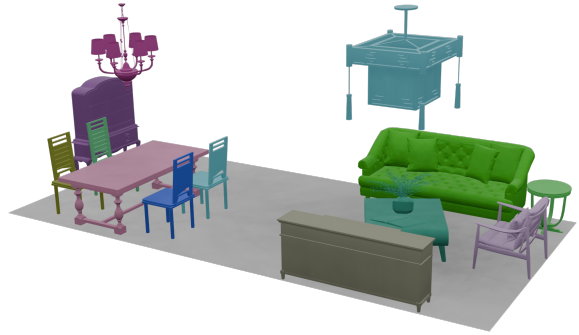
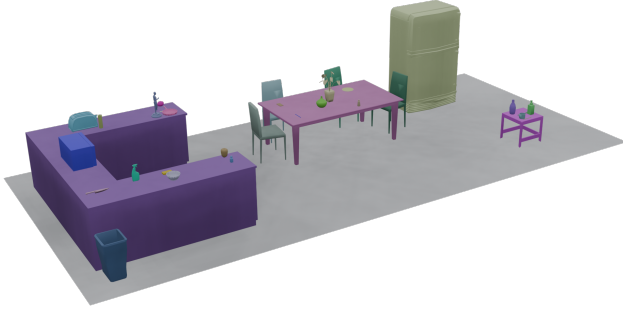


Figure 7. **Comparison of relational complexity between our pre-processed dataset (left) and the original 3D-FRONT dataset (right).** The sample from our dataset (left) contains a richer variety of indoor objects and exhibits complex, fine-grained spatial-physical relationships. In contrast, the sample from 3D-FRONT (right) lacks such fine-grained object interactions, particularly with small objects, highlighting the relational sparsity inherent in the original dataset that our preprocessing pipeline aims to address

sion in a scene.

**What is the effect of geometry-aware module in scene diffusion?** In this paper, we posit that geometry-aware diffusion is central to enabling generative models to understand and generate spatially and physically plausible scenes. Without the explicit integration of geometric information, conventional scene generation merely learns the statistical distribution of the dataset, a process inherently limited by the dataset’s scale and quality. However, the dataset distribution is insufficient to capture complex inter-object relationships and therefore fails to address object intersections within a scene. This relatively unexplored problem motivates us to integrate explicit awareness of the geometric state into the scene generation process at the current timestep. To evaluate our method’s performance without geometry feature diffusion, we remove the geometry-aware module. Consequently, the ablated model produces scenes with higher colliding rates and floating artifacts. This validates the effectiveness of geometry-aware module in achieving understanding and generating spatial and physically plausible results.

**What is the effect of perceiver-based transformer as geometry-aware module?** We advocate the use of perceiver-based transformer as our geometry-aware network. The architecture allows for processing large-scale point cloud inputs without structural modification, effectively reducing complexity of model without compromising performance. An alternative is to employ plain transformer. However, we observe it achieves equal performance at higher computational cost and deeper network structure.

## 10. Baselines

ATISS [32] generates indoor scenes by autoregressively adding objects. At each step, the model encodes the set of previously placed objects. It subsequently predicts the

attributes including 3D location, size, orientation, and semantic category.

**DiffuScene.** DiffuScene represents each scene as a fixed-size set of objects, where each object is characterized by a concatenation of properties including 3D location, size, orientation, semantic category, and geometric feature. During the generation, the model iteratively denoises a randomly initialized set of object attributes using a UNet-1D architecture with attention mechanisms. After denoising, object geometries are retrieved from a 3D database using the predicted shape codes and semantic labels.

**InstructScene.** InstructScene synthesizes 3D indoor scenes through a two-stage generative framework. In the first stage, the model learns a semantic graph prior capturing high-level object relationships and appearance features conditioned on natural language instructions. In the second stage, a layout decoder generates precise 3D layout attributes—including location, size, and orientation—for each object by denoising a noisy latent representation derived from the semantic graph. However, it relies on object retrieval from a predefined asset library, which inherently introduces limitations.

## 11. Additional results

In Fig.8, we provide additional qualitative results and zoom-in visualization, demonstrating our work’s capability to generate physically plausible and meaningful relational results.

## 12. User study

We conducted a perceptual user study to evaluate the physical and semantic plausibility of our method in comparison to ATISS, DiffuScene and InstructScene on the scene generation task. Fig.9 displays the scenes generated by the four methods alongside their corresponding post-physics simulation states. For each pair of results, participants were

asked to select the scene exhibiting greater physical consistency and scene rationality. We collected 57 valid responses and calculate the statistics. Our method secured 88.6% of the votes, demonstrating a substantial advantage over ATISS (0.9%), InstructScene (4.4%), and DiffuScene (6.1%).



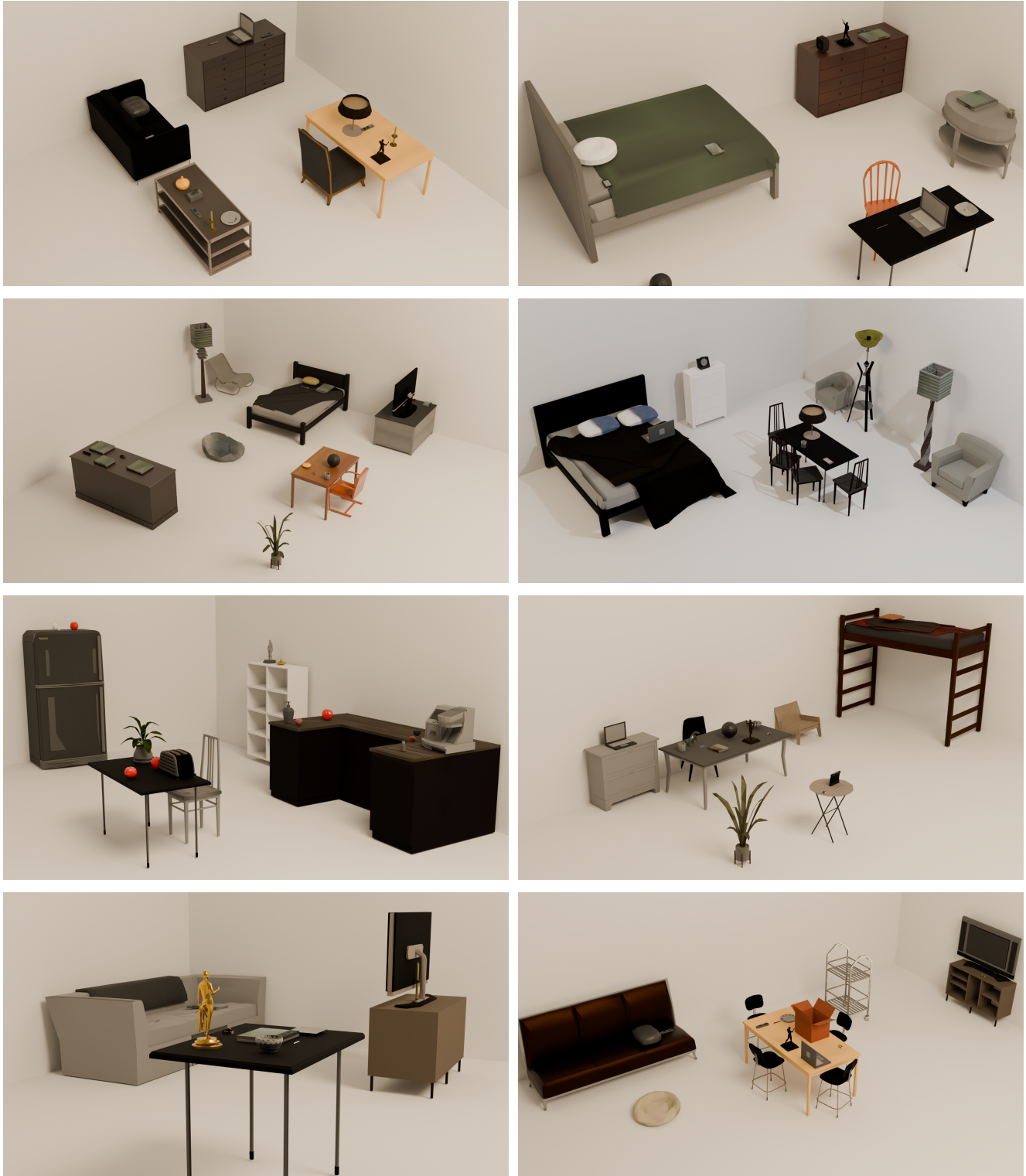
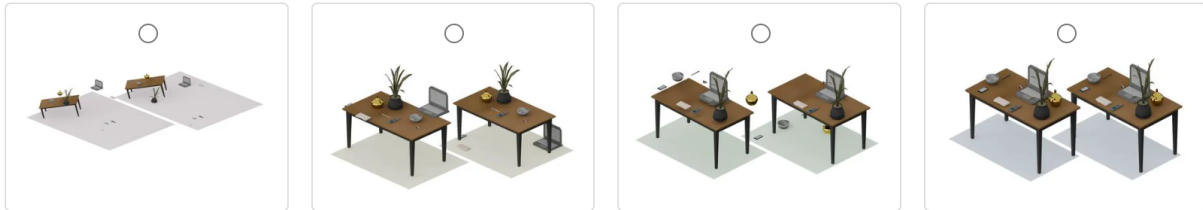


Figure 8. **Additional Qualitative results.** The gallery displays 8 randomly selected samples, demonstrating the diversity and physical plausibility of the generated 3D scenes.

1. In each of the following options, the image on the bottom left is the scene directly generated by the model, and the image on the top right is the scene obtained after physical simulation. Please select the option you think has the best physical consistency and scene plausibility.



2. In each of the following options, the image on the bottom left is the scene directly generated by the model, and the image on the top right is the scene obtained after physical simulation. Please select the option you think has the best physical consistency and scene plausibility.



3. In each of the following options, the image on the bottom left is the scene directly generated by the model, and the image on the top right is the scene obtained after physical simulation. Please select the option you think has the best physical consistency and scene plausibility.



Figure 9. **User Study UI.** Based on the generated scene and the scene after physical simulation, which were available for detailed inspection via a zoom function, participants were asked to choose the pair exhibiting greater physical consistency and scene rationality.