

Supplementary Materials: RelScene: A Benchmark and baseline for Spatial Relations in text-driven 3D Scene Generation

Anonymous Authors

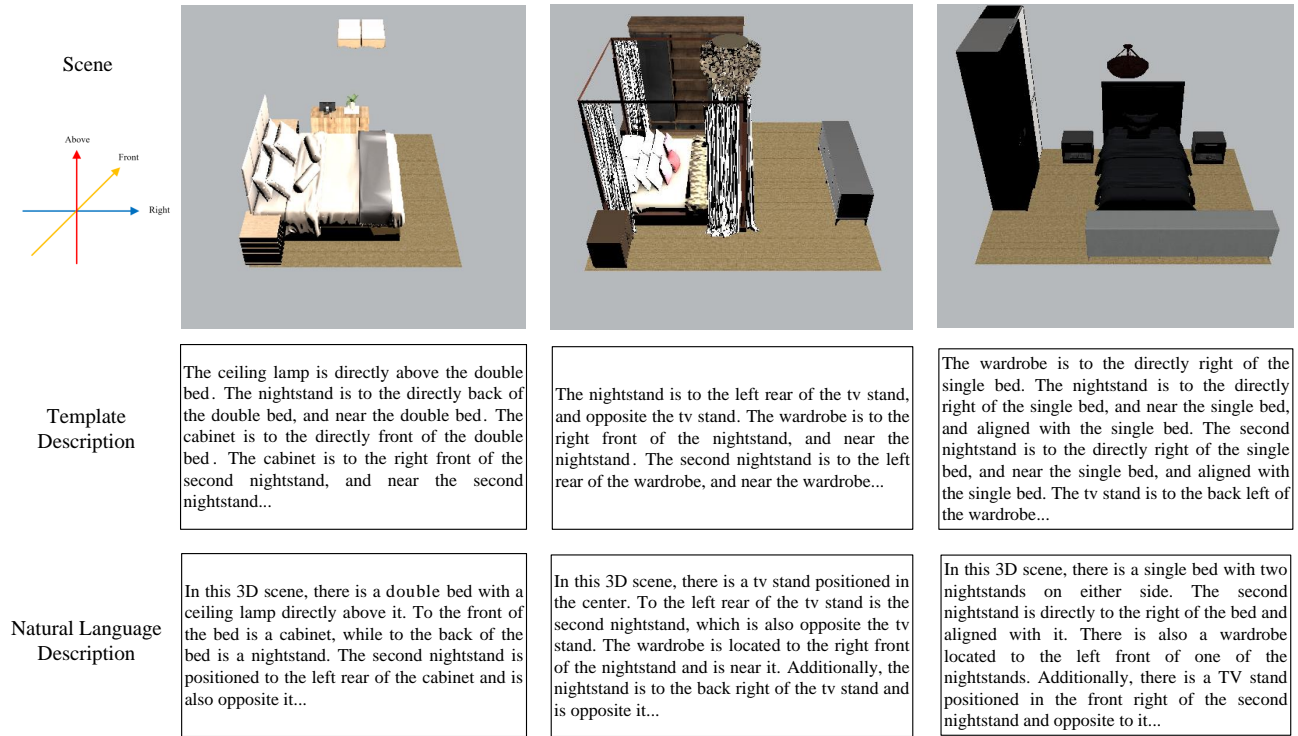


Figure 1: The samples of our proposed dataset and the text description.

1 DETAILS OF THE PROPOSED DATASET

Our new proposed dataset has scenes with multiple objects and textual descriptions for the generation task, which includes 13 types of relationships of the objects. Fig.1 shows the samples of our datasets.

1.1 Scene Data Processing

Our dataset is extended from 3D-FRONT dataset[3]. We select samples from the original dataset in three categories: bedrooms, living rooms, and dining rooms. In accordance with the methodologies outlined in recent studies[7][4], we have applied the identical dataset filtering procedures as employed in the ATISS framework. These procedures involved excluding scenes that exhibited excessive complexity, excessive simplicity, and the absence of typical object relationships. This process yielded 4041, 900, and 813 scenes in their respective subsets. In all these scenes, the number of objects ranged from 3 to 13, ensuring that the quantity of textual descriptions would not be excessive or insufficient. Each scene contains several types of information, including the object class, object positions, orientations, and object coordinates marked with eight

Table 1: Statistics of our Datasets.

Room Type	Scene	relation	Train	test	Valid
Bedroom	4041 rooms	24K pairs	3722	162	157
Dining Room	900 rooms	5.5K pairs	631	177	92
Living Room	813 rooms	4.6K pairs	543	192	78

points. We extract this information to generate template-based textual descriptions and provide essential data for natural language models.

1.2 Textual Description Generation

1.2.1 Template-based textual descriptions generation. The original 3D-FRONT dataset has no textual description for text-driven 3D scene generation.

To equip the 3D-Front dataset with text descriptions, we propose an algorithm for labeling fundamental spatial relationships by drawing insights from the extracted scene data. To label attributes and relationships between two objects, we employ triplets structured as

Table 2: Definition of basic relationships.

Relationship	Definition
Front Center	If the object is between the leftist and rightest bounds of another Object.
Side Center	If the object centroid is between the most forward and most behind bounds of another Object.
Vertical	If the object centroid is in both the front center and side center of another object.
High	If the z-value of the object centroid is bigger than another Object.
Front/Back	If the y-value of the object's centroid is bigger or smaller than another Object.
Left/Right	If the x-value of the object's centroid is bigger or smaller than another Object.
Aligned	The direction of the object equals to another Object.
Near	If the minimum Euclidean distance of the bounding box pair of two objects is smaller than the threshold.

Table 3: Definition of different relationships.

Keyword	Vertical	Frontal Center	Side Center	High	Front/Back	Left/Right	Aligned	Near
Near	×	-	-	Parallel	-	-	-	Y
Aligned with	×	-	-	Parallel	-	-	Aligned	-
Opposite	×	-	-	Parallel	-	-	Opposite	-
Back right	×	×	×	Parallel	Back	Right	-	-
Right front	×	×	×	Parallel	Front	Right	-	-
Left rear	×	×	×	Parallel	Back	Left	-	-
Front left	×	×	×	Parallel	Front	Left	-	-
Directly right	×	×	Y	Parallel	-	Right	-	-
Directly left	×	×	Y	Parallel	-	Left	-	-
Directly back	×	Y	×	Parallel	Back	-	-	-
Directly front	×	Y	×	Parallel	Front	-	-	-
Directly below	Y	-	-	Low	-	-	-	-
Directly above	Y	-	-	High	-	-	-	-

Table 4: The average length of the text description in the dataset.

		bedroom	living	dining
template	sentence	4.68	4.98	5.65
	word	89.25	111.27	110.41
natural language	sentence	5.07	5.65	4.98
	word	81.69	99.91	100.06

"subject-relationship-object." These basic relationships encompass attributes such as the volumes of objects A and B, the relative positioning of A in relation to B (e.g., above or below), and so forth. Tab.

2 represents all the basic relationships and the definition details. Then, we used these basic relationships to generate template-based textual descriptions among all the objects in the scene as initial text content. The initial text content contains C_n^2 (n is the number of objects in the scene) textual descriptions and is redundant. Hence, we applied a filtering algorithm based on probability to handle the previously acquired set of text descriptions. The filtering model considers the occurrences of typical relationships, volumes, and distances between objects to evaluate the importance of specific texts, which is defined as follows:

$$S(O_i, O_j) = W_{O_i} V_{O_i} + W_{O_j} V_{O_j} + D(O_i, O_j) + R(O_i, O_j) \quad (1)$$

where W_{O_i} , W_{O_j} is the select weight of the object, V_{O_i} , V_{O_j} is the volumes of the objects, D is the distance of two objects, and R is the relation weight of the two objects. Specifically, the initial select

weight of each object is set to 1 and is reduced if the object is selected. The V and the D are normalized into the range between 0 and 1. The $R(O_i, O_j) = 1 - P_r$, where the P_r is the proportion of the relation between O_i and O_j in all generated descriptions. The consideration of occurrences of typical relationships aims to balance the amount of different relationships to reduce the long-tailed distributions of the relationships. The consideration of the volumes of the objects can help the model focus on the main objects in the scene, and the distances can help select descriptions of the related objects.

After each filtering step, we reduce the coefficients of the selected objects for the next step, which can reduce too many repeating objects in the final text description. By employing this recursive calculation approach, we establish a mechanism that prioritizes selecting significant relationships and those involving the key objects (the objects with big volumes and more relationships) within the scene. Simultaneously, it maintains a certain likelihood of selecting less conspicuous relationships, thereby upholding the diversity of textual descriptions.

1.2.2 Natural language descriptions generation. To get closer to the text-driven 3D scene generation application scenario, we generate natural language descriptions with ChatGPT, a powerful language model for dialogue. To leverage its excellent natural language capabilities, we craft prompts *I give you a description of a 3D scene. Please summarize it in a paragraph, using up to 100 words. {template text descriptions}* to enable the model to grasp the spatial relationships between objects and reiterate the essence of template-based text descriptions using expressive and natural language. The ChatGPT can help to reduce the redundant words in the template text description. For example, the template text description "The King-size Bed is to the right front of the Dressing Table, and near the Dressing Table, and aligned with the Dressing Table" will be rewritten to the "In this 3D scene, there is a king-size bed positioned to the right front of a dressing table, aligned with it". The ChatGPT can also help to apply complex grammar to describe a scene, such as rewriting the "The Shelf is to the right front of the Wardrobe, and opposite the Wardrobe" to the "To the right front of the Wardrobe is a Shelf, which is also opposite the Wardrobe".

The rewritten description by the ChatGPT is more natural to the human inputs. It provides a more robust evaluation of the text-driven 3D scene generation models' capacity to extract and comprehend crucial information from natural language. Additionally, these descriptions offer a broader range of challenges and diversity for the task.

2 ADDITIONAL RESULTS

2.1 Text-conditioned scene generation with full training data

To compare the qualitative text-conditioned scene generation, we modify the ATISS to support the text conditioned as input. We adopt the Bert to encode the text into latent features as the condition information to input the ATISS. First, we compare our approach with the ATISS on MLA and MRA scores. From Tab.5, our approach improves the MLA scores from 0.293 to 0.337 and the MRA scores from 0.258 to 0.324 in the bedroom. This observation highlights a

Table 5: Performance comparison on Text-conditioned 3D scene generation with template description.

Metric	Method	Bedroom	Living	Dining
MLA(\uparrow)	ATISS*	0.293	0.095	0.084
	Our	0.337	0.209	0.204
MRA(\uparrow)	ATISS*	0.258	0.086	0.083
	Our	0.324	0.143	0.168
FID(\downarrow)	ATISS*	18.72	38.79	41.24
	Our	17.15	31.64	35.85
KID(\downarrow)	ATISS*	1.94	5.83	5.26
	Our	1.58	0.74	0.90
CKL(\downarrow)	ATISS*	0.82	0.66	0.72
	Our	0.34	0.25	0.24

Table 6: Performance comparison on Text-conditioned 3D scene generation with natural language description.

Metric	Method	Bedroom	Living	Dining
MLA(\uparrow)	ATISS*	0.254	0.092	0.082
	Our	0.316	0.184	0.154
MRA(\uparrow)	ATISS*	0.258	0.085	0.080
	Our	0.295	0.164	0.147
FID(\downarrow)	ATISS*	19.12	38.92	41.45
	Our	17.10	31.58	35.54
KID(\downarrow)	ATISS*	1.98	5.92	5.34
	Our	1.59	0.75	0.89
CKL(\downarrow)	ATISS*	0.85	0.71	0.77
	Our	0.33	0.26	0.26

Table 7: Performance comparison on 3D scene generation without constraints

Method	FID	KID
EpiGRAF[6]	107.2	102.3
VolumeGAN[9]	52.7	38.7
EG3D[1]	19.7	13.5
GIRAFFE[5]	56.5	46.8
GSN[2]	130.7	87.5
DisCOScene[8]	13.8	7.4
Ours	11.92	3.7

substantial improvement in both scores, signifying our approach’s proficiency in capturing the semantics of textual descriptions and adeptly controlling the spatial relationships among objects. Compared with adopting text features as condition information to input the model, the cross-attention in our approach is more effective in injecting the semantics into the text-conditioned generation tasks for the transformer model. Expanding upon previous research, we extend our evaluation to include a comparative analysis, employing FID, KID, and CKL scores to gauge the quality of the generated scenes in our approach and ATISS. As illustrated in Table 5, the performance trends based on FID, KID, and CKL scores consistently confirm that our approach outperforms ATISS. It is noted that the performance in the living room and the dining room is relatively poor compared to the bedroom. This is due to the training data of these two types of rooms is much less than the bedroom. Thus, it is necessary to consider the few-shot setting to apply the unannotated scene for the training.

A comparison with Tab.6 reveals that generating scenes under semantic constraints becomes more challenging with natural language descriptions, as evidenced by decreased MLA and MRA scores across all methods. This underscores the importance of addressing the challenges of free-form natural language in text-driven scene generation.

2.2 More comparison for no semantic constraints scene generation methods

We also consider other works focusing on building and modeling the 3D scene within the latent space, which can generate and render the scenes’ images with the camera’s parameters. From tab.7, compared to DisCoScene, we reduced the FID score from 13.8 to 11.92 and KID from 7.4 to 3.7. These methods mainly focus on how to model the scenes within the latent space. They hardly consider how to improve the layout of the objects in the scenes. Our proposed approach aims to generate the scenes with proper layout under the language constraint, which has the advantage of optimizing the scene layout to generate scenes.

REFERENCES

- [1] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. 2022. Efficient geometry-aware 3D generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16123–16133.
- [2] Terrance DeVries, Miguel Angel Bautista, Nitish Srivastava, Graham W Taylor, and Joshua M Susskind. 2021. Unconstrained scene generation with locally conditioned radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14304–14313.
- [3] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 2021. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10933–10942.
- [4] Jingyu Liu, Wenhan Xiong, Ian Jones, Yixin Nie, Anchit Gupta, and Barlas Oğuz. 2023. Clip-layout: Style-consistent indoor scene synthesis with semantic furniture embedding. *arXiv preprint arXiv:2303.03565* (2023).
- [5] Michael Niemeyer and Andreas Geiger. 2021. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11453–11464.
- [6] Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. 2022. EpigraF: Rethinking training of 3d gans. *Advances in Neural Information Processing Systems* 35 (2022), 24487–24501.
- [7] Jiapeng Tang, Yinyu Nie, Lev Markhasin, Angela Dai, Justus Thies, and Matthias Nießner. 2023. Diffuscene: Scene graph denoising diffusion probabilistic model for generative indoor scene synthesis. *arXiv preprint arXiv:2303.14207* (2023).
- [8] Yinghao Xu, Menglei Chai, Zifan Shi, Sida Peng, Ivan Skorokhodov, Aliaksandr Siarohin, Ceyuan Yang, Yujun Shen, Hsin-Ying Lee, Bolei Zhou, et al. 2023. DisCoScene: Spatially Disentangled Generative Radiance Fields for Controllable 3D-aware Scene Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4402–4412.
- [9] Yinghao Xu, Sida Peng, Ceyuan Yang, Yujun Shen, and Bolei Zhou. 2022. 3d-aware image synthesis via learning structural and textural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18430–18439.