MMGDreamer: Mixed-Modality Graph for Geometry-Controllable 3D Indoor Scene Generation

Zhifei Yang[1], Keyang Lu[2], Chao Zhang[3*], Jiaxing Qi[4], Hanqi Jiang[3], Ruifei Ma[3], Shenglin Yin[1], Yifan Xu[2], Mingzhe Xing[1], Zhen Xiao[1*], Jieyi Long[4], Xiangde Liu[3], Guangyao Zhai[5]

[1]Peking University [2]Beihang University [3]Beijing Digital Native Digital City Research Center [4]Theta Labs, Inc. [5]Technical University of Munich

Code Homepage

Introduction

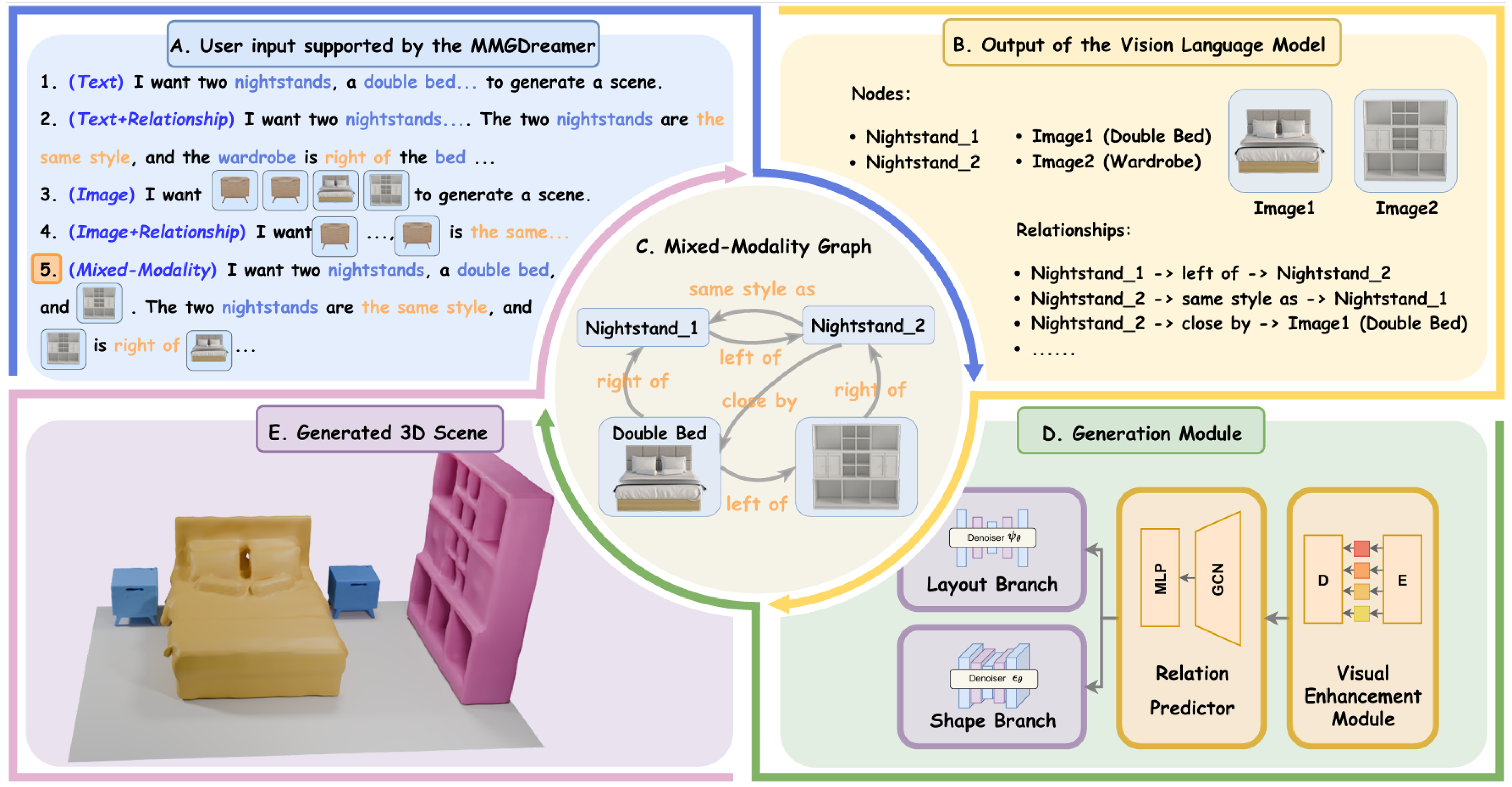

Figure 1: MMGDreamer processes a Mixed-Modality Graph to generate a 3D indoor scene, where object geometry can be precisely controlled. Starting from the fifth type of input (Mixed-Modality) shown in module A as an example, the framework utilizes a vision-language model (B) to produce a Mixed-Modality Graph (C). This graph is further refined by the Generation Module (D) to create a coherent and precise 3D scene (E).

Motivations:

- Current graph-based methods for indoor scene generation are constrained to text-based inputs and exhibit insufficient adaptability to flexible user inputs.
- The current indoor scene generation methods have poor geometric control of generated objects, and can not achieve accurate geometric control.
- Scene graphs serve as a powerful tool by succinctly abstracting the scene context and interrelations between objects, enabling intuitive scene manipulation and generation.

Contributions:

- We introduce a novel **Mixed-Modality Graph**, where nodes can selectively incorporate textual and visual modalities, allowing for precise control over the object geometry of the generated scenes and more effectively accommodating flexible user inputs.
- We present **MMGDreamer**, a dual-branch diffusion model for scene generation based on Mixed-Modality Graph, which incorporates two key modules: a visual enhancement module and a relation predictor, dedicated to construct node visual features and predict relations between nodes, respectively.
- Extensive experiments on the SG-FRONT dataset demonstrate that MMGDreamer attains higher fidelity and geometric controllability, and achieves state-of-the-art performance in scene synthesis, outperforming existing methods by a large margin.

Pipeline

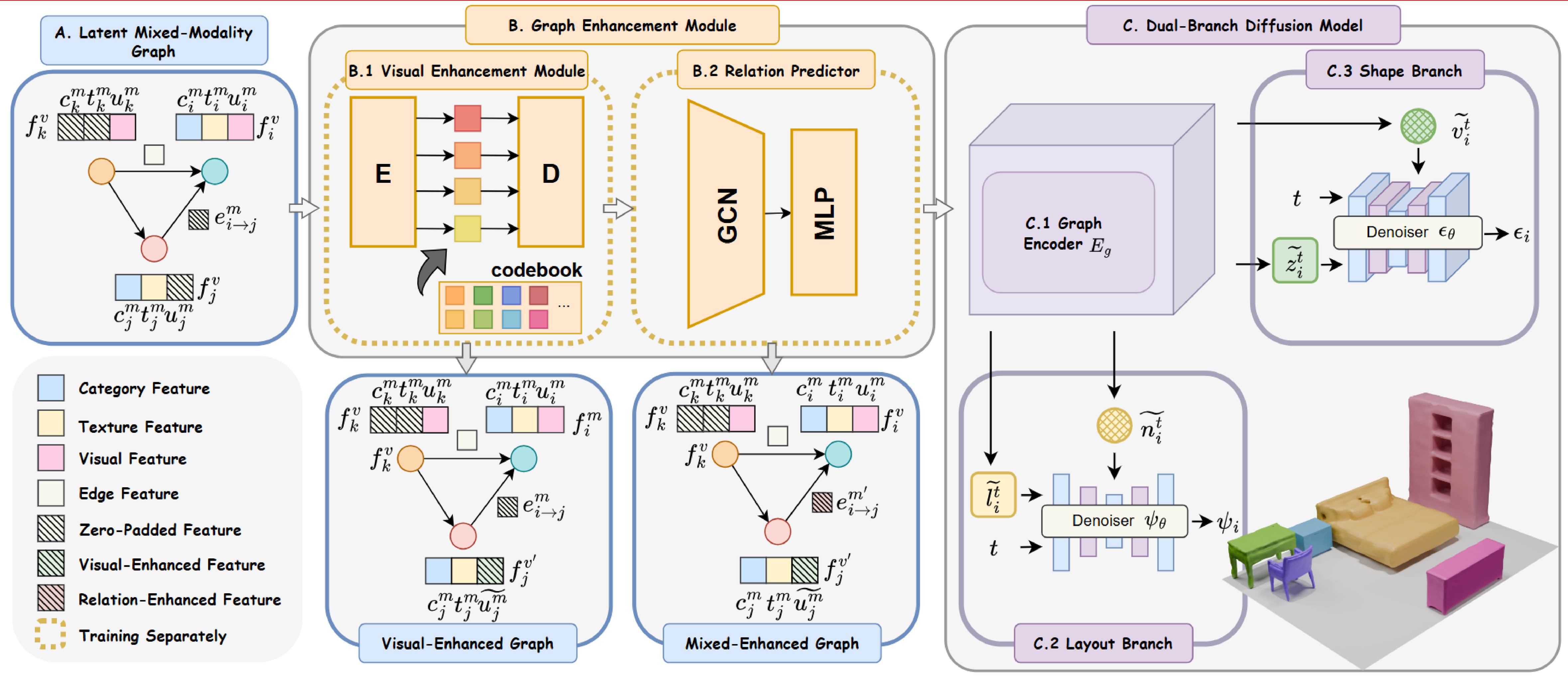

Figure 2: Overview of MMGDreamer. Our pipeline consists of the Latent Mixed-Modality Graph, the Graph Enhancement Module, and the Dual-Branch Diffusion Model. During inference, MMGDreamer initiates with the Latent Mixed-Modality Graph, which undergoes enhancement via the Visual Enhancement Module and the Relation Predictor, resulting in the formation of a Visual-Enhanced Graph and a Mixed-Enhanced Graph. The Mixed-Enhanced Graph is then input into the Graph Encoder E_g within the Dual-Branch Diffusion Model for relationship modeling, using a triplet-GCN structured module integrated with an echo mechanism. Subsequently, the Layout Branch (C.2) and the Shape Branch (C.3) use denoisers conditioned on the nodes' latent representations to generate layouts and shapes, respectively. The final output is a synthesized 3D indoor scene where the generated shapes are seamlessly integrated into the generated layouts.

Experimental Results

Method	Shape Representation	Bedroom FID	Bedroom $\mathrm{FID_{CLIP}}$	Bedroom KID	Living room FID	Living room $\mathrm{FID_{CLIP}}$	Living room KID	Dining room FID	Dining room $\mathrm{FID_{CLIP}}$	Dining room KID
Graph-to-3D (Dhamo et al. 2021)	DeepSDF (Park et al. 2019)	63.72	6.01	17.02	82.96	7.80	11.07	72.51	7.25	12.74
CommonLayout+SDFusion (Cheng et al. 2023)	txt2shape	68.08	5.61	18.64	85.38	7.23	10.04	64.02	6.92	5.08
EchoLayout+SDFusion (Cheng et al. 2023)	txt2shape	57.68	4.96	10.54	83.66	6.83	9.62	65.55	7.02	4.99
CommonScenes (Zhai et al. 2024c)	rel2shape	57.68	4.86	6.59	80.99	7.05	6.39	65.71	7.04	5.47
EchoScene (Zhai et al. 2024b)	echo2shape	48.85	4.26	1.77	75.95	6.73	0.60	62.85	6.28	1.72
MMGDreamer (MM+R)	echo2shape	45.75	3.84	1.72	68.94	6.19	0.40	55.17	5.86	0.05

Table 1: **Scene generation realism** is quantified by comparing generated top-down renderings with real scene renderings at a resolution of 256^2 pixels, using FID, $\mathrm{FID_{CLIP}}$ and KID. The best and second results are highlighted.

Method	Metric	Bed	N.stand	Ward.	Chair	Table	Cabinet	Lamp	Shelf	Sofa	TV stand
Graph-to-3D (Dhamo et al. 2021)	MMD (↓)	1.56	3.91	1.66	2.68	5.77	3.67	6.53	6.66	1.30	1.08
CommonScenes (Zhai et al. 2024c)		0.49	0.92	0.54	0.99	1.91	0.96	1.50	2.73	0.57	0.29
EchoScene (Zhai et al. 2024b)		0.37	0.75	0.39	0.62	1.47	0.83	0.66	2.52	0.48	0.35
MMGDreamer (I+R)		0.22	0.41	0.24	0.35	0.55	0.71	0.34	1.58	0.43	0.24
Graph-to-3D (Dhamo et al. 2021)	COV (%, ↑)	4.32	1.42	5.04	6.90	6.03	3.45	2.59	13.33	0.86	1.86
CommonScenes (Zhai et al. 2024c)		24.07	24.17	26.62	26.72	40.52	28.45	36.21	40.00	28.45	33.62
EchoScene (Zhai et al. 2024b)		39.51	25.59	37.07	17.25	35.05	43.21	33.33	50.00	41.94	40.70
MMGDreamer (I+R)		42.59	30.81	44.44	19.95	44.12	49.38	40.56	70.00	47.31	45.35
Graph-to-3D (Dhamo et al. 2021)	1-NNA (%, ↓)	98.15	99.76	98.20	97.84	98.28	98.71	99.14	93.33	99.14	99.57
CommonScenes (Zhai et al. 2024c)		85.49	95.26	88.13	86.21	75.00	80.17	71.55	66.67	85.34	78.88
EchoScene (Zhai et al. 2024b)		72.84	91.00	81.90	92.67	75.74	69.14	78.90	35.00	69.35	78.49
MMGDreamer (I+R)		69.44	90.52	74.81	89.56	68.85	68.35	72.38	30.00	62.37	73.26

Table 2: **Object-level generate-on performance.** We present MMD, COV, and 1-NNA metrics to assess the quality and diversity of the generated shapes. **I** represents nodes using image representations. **R** denotes the relationships of nodes.

Visualization Results

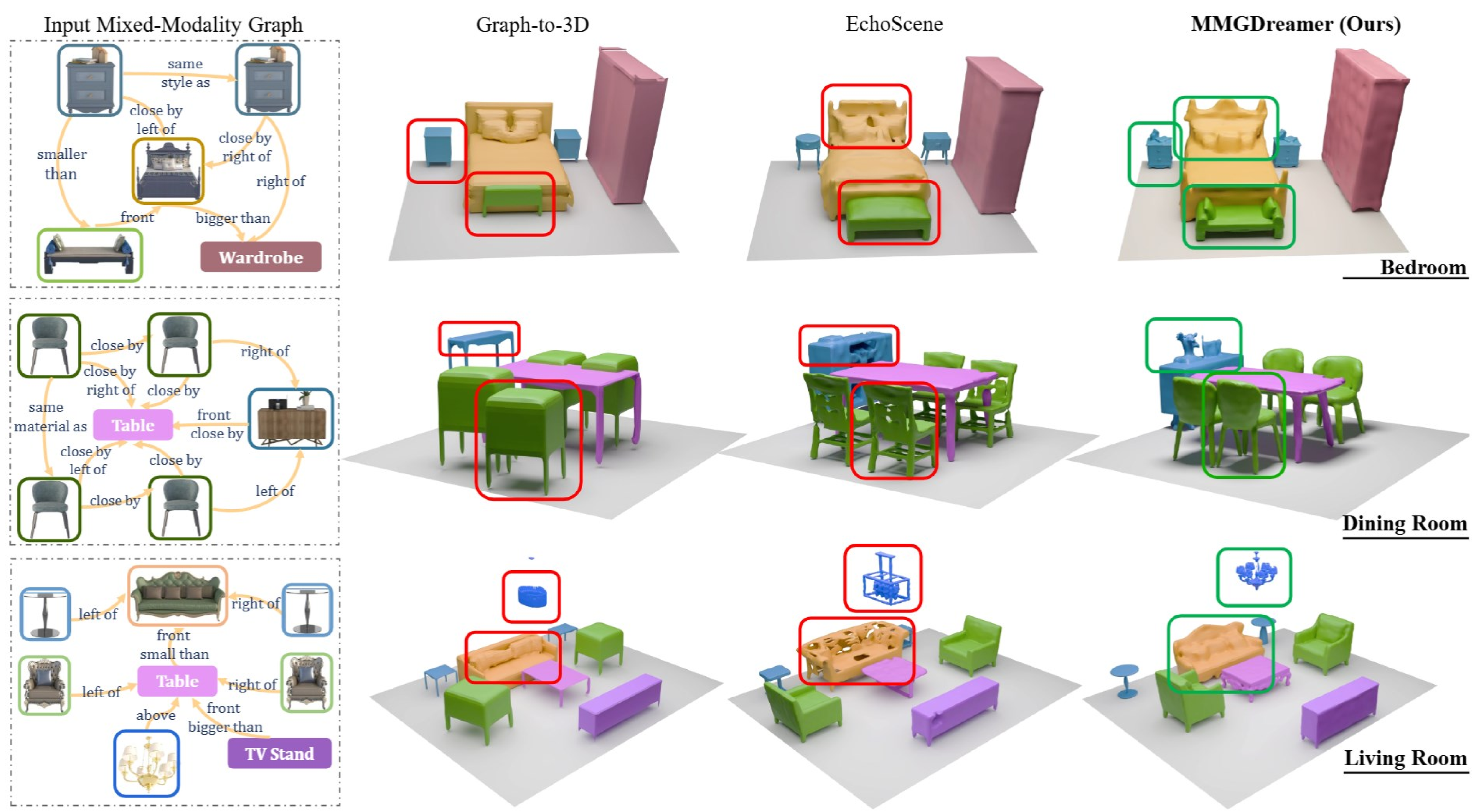

Figure 3: Qualitative comparison with other methods. The first column shows the input mixed-modality graph, which visualizes only the most critical edges in the scene. Red rectangles denote areas of inconsistency in the generated scenes, while green rectangles signify regions of consistent generation.

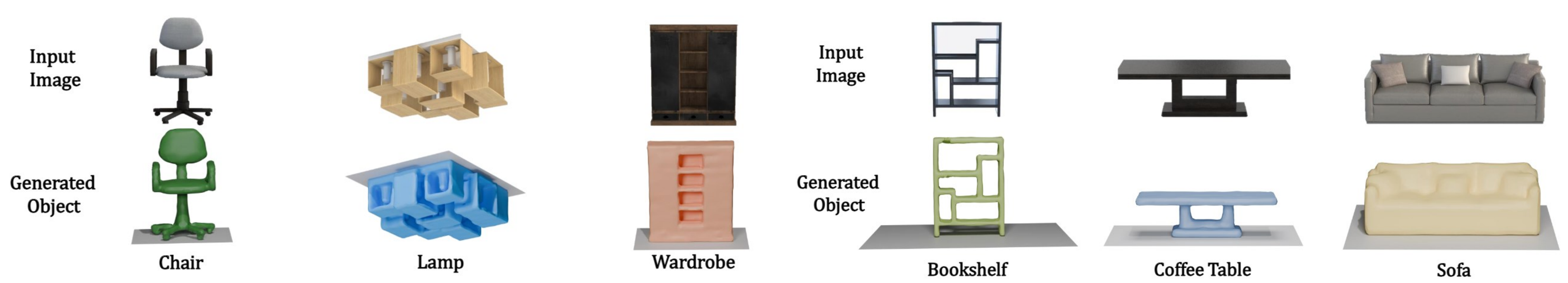

Figure 4: Qualitative results on object generation. The top row shows the input images of various furniture items, the middle row displays the corresponding generated objects in the scenes, and the bottom row provides the object categories.

Limitations and Future Work

While our method successfully integrates visual information, we have intentionally focused on generating objects with accurate geometric shapes and coherent scene layouts, deliberately excluding texture and material details for simplicity and control. We recognize that including texture and material information presents an exciting opportunity for future work. By enhancing the method to better leverage visual data, we plan to generate scenes with richer texture details.

Conclusion

We present MMGDreamer, a dual-branch diffusion model for geometry-controllable 3D indoor scene generation, leveraging a novel Mixed-Modality Graph that integrates both textual and visual modalities. Our approach, enhanced by a Visual Enhancement Module and a Relation Predictor, provides precise control over object geometry and ensures coherent scene layouts.