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[4], Mingzhe Xing[1], Zhen Xiao[1†], Jieyi Long[5], Xiangde Liu[3], and Guangyao Zhai[6]

[1] School of Computer Science, Peking University [2] School of Artificial Intelligence, Beihang University
[3] Beijing Digital Native Digital City Research Center [4] School of Computer Science and Engineering, Beihang University
[5] Theta Labs, Inc. [6] Technical University of Munich

yangzhifei@stu.pku.edu.cn, ariczhang2009@gmail.com, xiaozhen@pku.edu.cn, guangyao.zhai@tum.de

Abstract

Controllable 3D scene generation has extensive applications in virtual reality and interior design, where the generated scenes should exhibit high levels of realism and controllability in terms of geometry. Scene graphs provide a suitable data representation that facilitates these applications. However, current graph-based methods for scene generation are constrained to text-based inputs and exhibit insufficient adaptability to flexible user inputs, hindering the ability to precisely control object geometry. To address this issue, we propose **MMGDreamer**, a dual-branch diffusion model for scene generation that incorporates a novel **Mixed-Modality Graph**, visual enhancement module, and relation predictor. The mixed-modality graph allows object nodes to integrate textual and visual modalities, with optional relationships between nodes. It enhances adaptability to flexible user inputs and enables meticulous control over the geometry of objects in the generated scenes. The visual enhancement module enriches the visual fidelity of text-only nodes by constructing visual representations using text embeddings. Furthermore, our relation predictor leverages node representations to infer absent relationships between nodes, resulting in more coherent scene layouts. Extensive experimental results demonstrate that MMGDreamer exhibits superior control of object geometry, achieving state-of-the-art scene generation performance.

Project page —
https://yangzhifeio.github.io/project/MMGDreamer

Introduction

Deep generative models have initiated a new era of artificial intelligence-generated content, driving developments in natural language generation (Zheng et al. 2024), video synthesis (Liao et al. 2024a), and 3D generation (Poole et al. 2022). Controllable Scene Generation refers to generating realistic 3D scenes based on input prompts, allowing for precise control and adjustment of specific objects within those scenes. It is widely applied in Virtual Reality (Bautista et al. 2022), Interior Design (Çelen et al. 2024), and Embodied Intelligence(Yang et al. 2024; Zhai et al. 2024a), providing immersive experiences and enhancing decision-making processes. Within these applications, scene graphs serve as a powerful tool by succinctly abstracting the scene context and interrelations between objects, enabling intuitive scene manipulation and generation (Dhamo et al. 2021).

Despite the significant progress made by retrieval-based (Lin and Mu 2024), semi-generative (Ren et al. 2024), and fully-generative (Zhai et al. 2024c) methods in graph-based controllable scene generation, these approaches predominantly rely on textual descriptions to construct input scene graphs. While text serves as a high-level representation encapsulating rich semantic information, it falls short in accurately describing the geometry of objects in the generated scenes, resulting in inadequate geometric control over the generated objects (Rombach et al. 2022). Moreover, each node in the scene graph contains only textual information about object category, which limits its adaptability to flexible user input. To address these limitations, we introduce MMGDreamer, a dual-branch diffusion model designed for processing multimodal information, incorporating a novel Mixed-Modality Graph (MMG) as a key component. As depicted in Fig. 1, the node of MMG can be represented in three ways: text, image, or a combination of both. Additionally, edges between nodes can be selectively provided or omitted based on user input. This flexible graph structure supports five types of user input, as illustrated in Fig. 1.A, significantly enhancing adaptability to diverse user demands and enabling precise control over object geometry in generated scenes.

To fully leverage the capabilities of MMG, MMGDreamer features two pivotal modules: the visual enhancement module and the relation predictor. When nodes of the input scene graph contain solely textual information, the visual enhancement module employs text embeddings to construct visual representations of these nodes. By incorporating visual priors associated with the text, this approach enriches the visual content of nodes, enhancing geometric control over the generated objects. The relation predictor, a relationship classifier based on the GCN, leverages prior knowledge and node representations within the scene to in-

[*]Work done at Beijing Digital Native Digital City Research Center.
[†]Corresponding author.
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

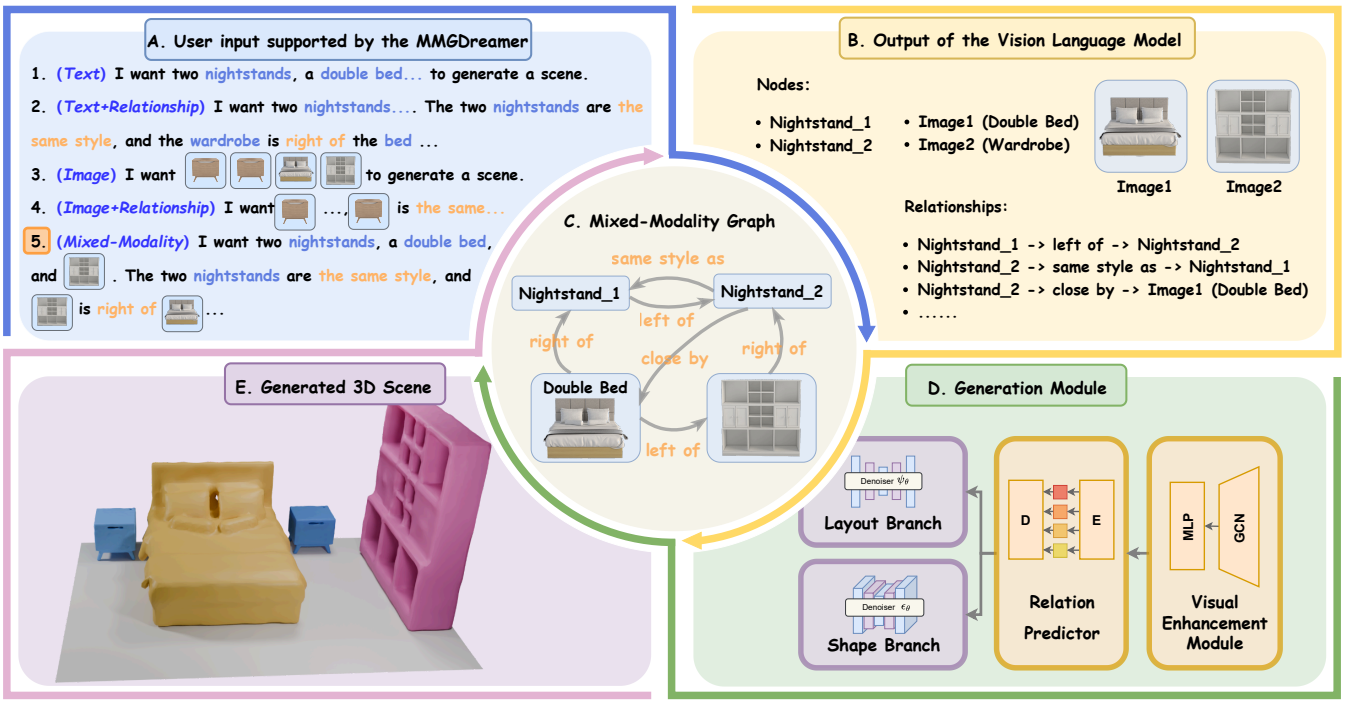

Figure 1: **MMGDreamer** processes a Mixed-Modality Graph to generate a 3D indoor scene, where object geometry can be precisely controlled. Starting from the fifth type of input (Mixed-Modality) shown in module A as an example, the framework utilizes a vision-language model (B) to produce a Mixed-Modality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

fer relationships between nodes in the absence of explicit relational information. By capturing global and local scene-object relationships, this module ensures the generation of more coherent and contextually appropriate scene layouts. We briefly summarize our primary contributions as follows:

- 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.

Related Work

Scene Graph. Scene graph provides a structured and hierarchical representation of complex scenes by using nodes (objects) and edges (relationships) (Zhou, While, and Kalogerakis 2019). Following their introduction, subsequent works have refined hierarchical scene graph (Rosinol et al. 2020) and focused on predicting local inter-object relationships (Koch et al. 2024; Liao et al. 2024b). Such advancements have driven the widespread application of scene graphs across both 2D and 3D domains, enabling sophisticated tasks such as image synthesis (Johnson, Gupta, and Fei-Fei 2018; Wu, Wei, and Lin 2023) and caption generation (Basioti et al. 2024) in 2D, as well as video synthesis (Cong et al. 2023), 3D scene understanding (Wald et al. 2020) and scene synthesis (Para et al. 2021) in 3D. However, in the current 3D indoor scene synthesis tasks, scene graphs are predominantly derived from text input by users (Strader et al. 2024). We propose the MMG, which more effectively accommodates flexible user inputs.

3D Scene Generation. 3D scene generation is an area of ongoing research that focuses on developing plausible layouts (Engelmann et al. 2021) and generating accurate object shapes (Xu et al. 2023). A substantial body of contemporary research synthesizes scenes from text (Fang et al. 2023), panorama (Wang et al. 2023; Hara and Harada 2024), or spatial layout (Yan et al. 2024; Jyothi et al. 2019) using autoregressive paradigms (Wang, Yeshwanth, and Nießner 2021), object decoupling techniques (Zhang et al. 2024; Epstein et al. 2024), or prior learning (Höllein et al. 2023). In particular, CommonScenes (Zhai et al. 2024c) utilizes scene graphs as conditions and adopts an end-to-end framework to generate both object shapes and scene layouts simultaneously. EchoScene (Zhai et al. 2024b) advances the Common-

Scenes by incorporating an information echo scheme. Nevertheless, these approaches are inadequate for effectively controlling object geometry. To overcome this limitation, we introduce MMGDreamer, which fully leverages the MMG to achieve precise control over object geometry.

Preliminary

Scene Graph Representation

Scene graph can be formally defined as a directed graph $\mathcal{G} = \{\mathcal{V}, \mathcal{R}\}$, where $\mathcal{V} = \{v_i \mid i \in \{1, \ldots, N\}\}$ represents the set of nodes (objects in the scene) and $\mathcal{R} = \{r_{i \to j} \mid i, j \in \{1, \ldots, N\}, i \neq j\}$ represents the set of edges (relationships between objects). The node v_i represents an object in the scene and contains information about the object's category. The edges $r_{i \to j}$ define the connections between objects, which can include spatial relationships (e.g., front/behind) or attribute relationships (e.g., same style).

Scene Graph Encoder

Graph Convolutional Network (GCN) (Johnson, Gupta, and Fei-Fei 2018) facilitates the processing of graph-structured data by learning node representations through the layer-by-layer aggregation of features from neighboring nodes. In our work, we utilize the Triplet Graph Convolutional Network (Triplet-GCN) as scene graph encoder to process the scene graph, assuming that the initial node and edge attribute features are given by $(\delta_{v_i}^{(0)}, \delta_{r_{i \to j}}^{(0)}, \delta_{v_j}^{(0)})$. Specifically, each layer of GCN updates the node and edge representations according to the following formula:

$$(\gamma_{v_i}^{l}, \delta_{r_{i \to j}}^{l+1}, \gamma_{v_j}^{l}) = \text{MLP}_1(\delta_{v_i}^{l}, \delta_{r_{i \to j}}^{l}, \delta_{v_j}^{l}), \tag{1}$$

$$\delta_{v_i}^{l+1} = \gamma_{v_i}^{l} + \text{MLP}_2\left(\text{Avg}(\gamma_{v_j}^{l} \mid v_j \in \mathcal{N}_G(v_i))\right), \tag{2}$$

where $l \in \{1, \ldots, L-1\}$ denotes an independent layer within the Triplet-GCN, $\mathcal{N}_G(v_i)$ represents the set of all neighboring nodes of v_i, Avg indicates the use of average pooling operation, and MLP_1 and MLP_2 refer to the Multi-Layer Perceptron (MLP) layers.

Latent Diffusion Model

Latent Diffusion Model (LDM) (Rombach et al. 2022) generally involves two Markov processes: a forward process that incrementally corrupts the data and a reverse process that progressively denoises it. Given a sample x_0, the LDM first employs a pre-trained VQ-VAE (Van Den Oord, Vinyals et al. 2017) to encode x_0 into a reduced-dimensional latent representation z_0. The forward process is defined by:

$$q(z_t \mid z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t} z_{t-1}, \beta_t \mathbf{I}), \tag{3}$$

where t ranges from 1 to T, with T denoting the total number of timesteps. The parameter β_t controls the noise level introduced at each timestep t. In the reverse process, z_t is processed through a denoiser ϵ_θ, such as UNet (Ronneberger, Fischer, and Brox 2015), to estimate the noise, enabling a progressive denoising process to recover a clean latent representation. The objective function can be written as:

$$\mathcal{L}_{LDM} = \mathbb{E}_{z_t, t, \epsilon \sim \mathcal{N}(0,1)}\left[\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2\right], \tag{4}$$

where c is a condition to guide the reverse process.

Method

We propose MMGDreamer, a framework adept at handling MMG as input for indoor scene synthesis tasks, as illustrated in Fig. 2. The MMG is a novel graph structure where nodes can optionally carry textual or visual information, thereby more effectively accommodating flexible user inputs. MMGDreamer first utilizes CLIP and an embedding layer to encode the MMG, producing the Latent Mixed-Modality Graph (LMMG). We then apply the visual enhancement module to construct visual information in the nodes of the LMMG, yielding a Visual-Enhanced Graph. Subsequently, a Relation Predictor is utilized to predict the missing edges between nodes, forming the Mix-Enhanced Graph. Finally, we model the relations within the scene using the Graph Encoder and employ a dual-branch diffusion model to generate the corresponding layout and shape, synthesizing the 3D indoor scene.

Mixed-Modality Graph

Generating fine-grained scenes using only text information is insufficient, as it cannot precisely control the geometry of generated objects. At the same time, users' flexible input should be multimodal, allowing for the selective input of text or images based on specific needs, as shown in Fig. 1.A. However, existing methods (Hu et al. 2024) do not support this input format. Graphs, as a compact and flexible structural representation, enable the effective encoding of diverse attributes within nodes, facilitating the seamless integration of multimodal information. Furthermore, users' textual descriptions often lack information about the relationships between all objects. While methods such as EchoScene(Zhai et al. 2024b) and CommoScenes (Zhai et al. 2024c) utilize graphs to generate scenes, they impose strict relation constraints, making them less user-friendly. A graph structure that mimics natural language should feature sparse edge relations. To address these issues, we propose the Mixed-Modality Graph, a novel graph where nodes can contain both textual and visual modalities, and edges are selectable.

A Mixed-Modality Graph $\mathcal{G}^m = \{\mathcal{V}^m, \mathcal{R}^m\}$ contains nodes and their relations:

$$\mathcal{V}^m = \{v_i^m \mid i \in 1, \ldots, N\}, \tag{5}$$

$$\mathcal{R}^m = \{r_{i \to j}^m \mid i, j \in \{1, \ldots, N\}, i \neq j\}. \tag{6}$$

Each node $v_i^m = \{[o_i^m, i_i^m] \mid [o_i^m] \mid [i_i^m]\}$ represents an object with text category information $[o_i^m]$, image information $[i_i^m]$ or both text category and image information $[o_i^m, i_i^m]$, as shown in Fig. 1.C. Despite the fact that MMG is generally easier to obtain than 3D spatial layouts (Schult et al. 2024), we have also devised a text prompt to query the vision language model such as GPT-4V (Achiam et al. 2023), enabling the parsing of MMG from unstructured text and image inputs, as illustrated in Fig. 1.B.

Assuming $v_i^m = [o_i^m, i_i^m]$, we utilize embedding layers to encode the category information o_i^m and the relational information of edge $r_{i \to j}^m$, transforming them into c_i^m and $e_{i \to j}^m$, respectively. To enrich high-level semantic features while simultaneously encoding image information, we leverage the

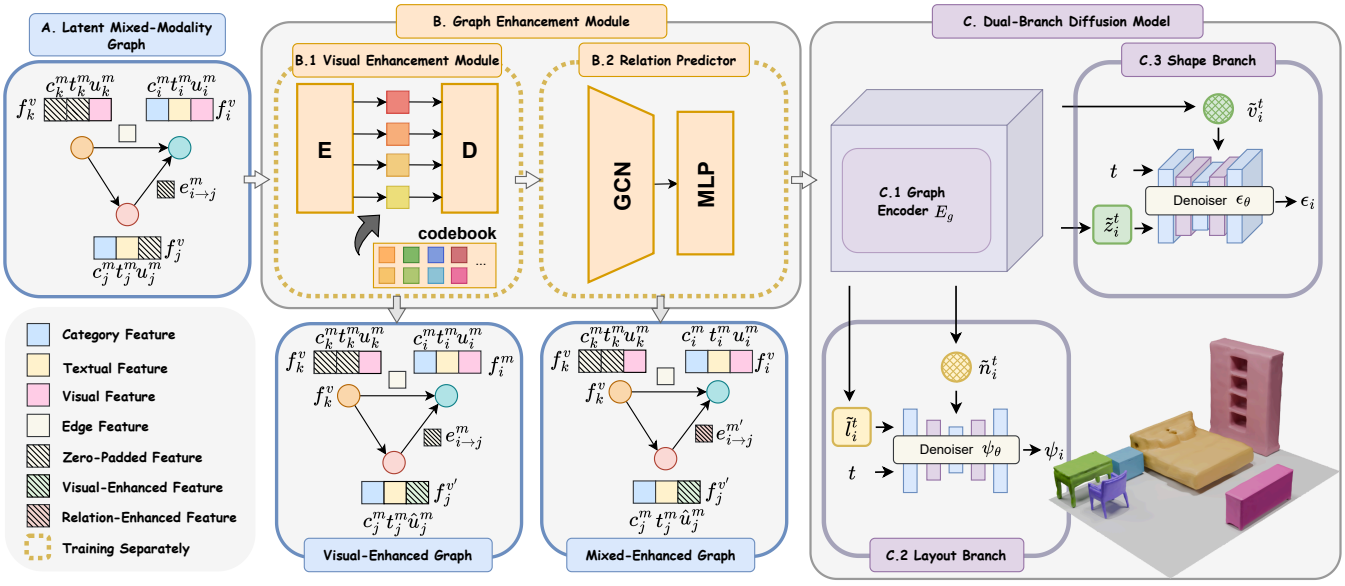

Figure 2: **Overview of MMGDreamer.** Our pipeline consists of the Latent Mixed-Modality Graph, the Graph Enhancement Module, and the Dual-Branch Diffusion Model. During inference, MMGDreamer initiates with the Latent Mixed-Modality Graph, which undergoes enhancement via the Visual Enhancement Module 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.

pre-trained and frozen visual language model CLIP (Radford et al. 2021), using its text encoder to transform o_i^m into t_i^m and its image encoder to convert i_i^m into u_i^m. To ensure consistent processing within MMGDreamer, we apply zero-padding at the feature level for missing node modality information or edge relationships between two nodes. As illustrated in Fig. 2.A, a LMMG $\mathcal{G}_l^m$ can be uniformly represented as:

$$\mathcal{F}_{\mathcal{V}^m} = \{f_i^v = [c_i^m, t_i^m, u_i^m] \mid i \in \{1, \ldots, N\}\}, \quad (7)$$

$$\mathcal{F}_{\mathcal{R}^m} = \left\{f_{i\to j}^e = [e_{i\to j}^m] \mid i, j \in \{1, \ldots, N\}, i \neq j\right\}, \quad (8)$$

where $\mathcal{F}_{\mathcal{V}^m}$ represents the set of node features, and $\mathcal{F}_{\mathcal{R}^m}$ represents the set of edge features.

Visual Enhancement Module

Incorporating visual features within graph nodes enhances the generation of object geometry. However, in the LMMG, some nodes only contain textual information. We introduce a visual enhancement module to bolster the ability to generate object shapes. This module employs an architecture similar to VQ-VAE, comprising an encoder E, a decoder D, and a codebook $\mathcal{C}$, to effectively construct visual features from the textual features of nodes within the LMMG. The encoder E processes the textual features t_i^m into latent vectors $h_i^m = E(t_i^m)$. These latent vectors are then quantized using the codebook $\mathcal{C}$, which contains a set of embedding vectors $\{e_k\}_{k=1}^K$. The quantization process selects the n nearest embedding vectors from the codebook:

$$\hat{h}_i^m = \{e_{k_j} \mid k_j \in \arg\min_{e_{k_l} \in \mathcal{C}} \sum_{l=1}^{n} \|h_i^m - e_{k_l}\|^2\}. \quad (9)$$

The quantized latent vectors $\hat{h}_i^m$ are subsequently processed by the decoder D to generate visual features $\hat{u}_i^m = D(\hat{h}_i^m)$. The training objective for the visual enhancement module is to maximize the evidence lower bound (ELBO) for the likelihood of the data:

$$\mathcal{L}_r = \mathbb{E}_{h\sim q_E(h|t)}\left[\log p_D(u|h) - \beta D_{\mathrm{KL}}(q_E(h|t)\|p(h))\right], \quad (10)$$

where $q_E(h|t)$ denotes the latent vector distribution given the textual features, $p_D(u|h)$ is the likelihood of the visual features given the latent vectors, and D_{KL} denotes the Kullback-Leibler divergence. The prior $p(h)$ is typically a Gaussian distribution, and β is a weighting factor. To address the non-differentiable nature of the quantization process, the Gumbel-Softmax relaxation (Jang, Gu, and Poole 2016) technique is applied to optimize the ELBO. Utilizing this VQ-VAE-based framework, the visual enhancement module produces a Visual-Enhanced Graph $\mathcal{G}_I^m$, enhancing the capability of the LMMG to generate accurate and detailed object geometry for scene generation tasks.

Relation Predictor

Relations are crucial in indoor scene generation, as they impact layout configuration. To address the challenge of missing relationships among nodes in the LMMG, we develop

a Relation Predictor that infers these connections, enabling the generation of more reasonable layouts. The Relation Predictor takes triples of latent representations $(f_i^v, f_{i\to j}^e, f_j^v)$ as input. In cases where relationships are missing, $f_{i\to j}^e$ is filled with zeros to ensure consistency in the feature space. The module comprises a GCN layer followed by a series of MLP layers. The GCN layer processes the input triples to capture the relational context between nodes, while the MLP layers further refine the edge predictions. The Relation Predictor is trained using a cross-entropy loss, defined as:

$$\mathcal{L}_e = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic}\log(\hat{y}_{ic}) \tag{11}$$

where N is the number of node pairs, C is the number of edge classes, y_{ic} is the one-hot encoded true label, and $\hat{y}_{ic}$ is the predicted probability. The Relation Predictor refines the graph $\mathcal{G}_I^m$ into a Mixed-Enhanced Graph $\mathcal{G}_E^m$, by predicting and integrating missing node relationships to improve overall layout coherence.

Shape and Layout Branch

We employ a dual-branch diffusion model to generate object shapes and scene layouts. To facilitate effective information exchange and relationship modeling between nodes during each denoising process, as depicted in Fig. 2.C.1, we employ a triplet-GCN structured module that integrates the echo mechanism (Zhai et al. 2024b) as a Graph Encoder E_g.

Shape Branch. For the shape branch, as shown in Fig. 2.C.2, we use Truncated Signed Distance Field (Curless and Levoy 1996) as shape representations and employ a pretrained and frozen VQ-VAE to encode them into latent representations z_i^0 and decode them back. At each denoising step $t \in \{1, 2, \ldots, T\}$, E_g is applied to process the latent codes z_i^t and the latent graph $\mathcal{G}_t^l$ (which originates from $\mathcal{G}_E^m$), yielding updated representations $\tilde{z}_i^t$ and $\tilde{\mathcal{G}}_t^l$. The updated nodes of $\tilde{\mathcal{G}}_t^l$, denoted as $\tilde{\mathcal{V}}_t = \{\tilde{v}_i^t\}$, are used as conditions for denoiser ϵ_θ (3D-UNet). The training objective is to minimize the deviation between the true noise ϵ and the predicted noise $\epsilon_\theta(\tilde{z}_i^t, t, \tilde{v}_i^t)$. The loss function is defined as:

$$\mathcal{L}_s = \mathbb{E}_{\tilde{z}_i^t, \epsilon\sim\mathcal{N}(0,1),t}\left\|\epsilon - \epsilon_\theta(\tilde{z}_i^t, t, \tilde{v}_i^t)\right\|^2. \tag{12}$$

Layout Branch. We utilize object bounding boxes to represent the layout of the scene. Each bounding box l_i^0 is characterized by its location $\mathbf{t}_i^0 \in \mathbb{R}^3$, size $\mathbf{s}_i^0 \in \mathbb{R}^3$, and rotation angle φ_i^0. Specifically, the rotation angle φ_i^0 is parameterized by $[\cos(\varphi_i^0), \sin(\varphi_i^0)]^\top$. To ensure proper scale and numerical stability during training, $\mathbf{t}_i^0$ and $\mathbf{s}_i^0$ are normalized. As illustrated in Fig. 2.C, the layout branch utilizes E_g for relationship modeling. This results in updated latent layout representations $\tilde{l}_i^t$ and refined graph node embeddings $\tilde{\mathcal{N}}_t = \{\tilde{n}_i^t\}$. Conditioned on the updated node embeddings, a 1D-UNet is utilized as the denoiser ψ_θ for the denoising process. The corresponding loss function is formulated as:

$$\mathcal{L}_l = \mathbb{E}_{\tilde{l}_i^t, \psi\sim\mathcal{N}(0,1),t}\left\|\psi - \psi_\theta(\tilde{l}_i^t, t, \tilde{n}_i^t)\right\|^2. \tag{13}$$

The overall training objective for the layout and shape branches is expressed as:

$$\mathcal{L}_o = \alpha_1\mathcal{L}_l + \alpha_2\mathcal{L}_s, \tag{14}$$

where α_1 and α_2 are weighting factors.

Training and Inference Strategy

The training process is divided into two stages. In the first stage, the visual enhancement module is trained with the loss function $\mathcal{L}_r$, which utilizes textual information from nodes to construct the corresponding visual features. The Relation Predictor is trained with $\mathcal{L}_e$ using triplet representations of the graph. In the second stage, the LMMG serves as the input, and the loss function $\mathcal{L}_o$ is employed to jointly optimize the graph encoder with the layout and shape branches, as depicted in Fig. 2.A and C. During inference, as shown in Fig. 2, the LMMG is processed through modules B and C to generate the indoor scene. Please see the Supplementary Material for further details.

Experiments

Experimental Settings

Evaluation Dataset. We validate our approach using the SG-FRONT dataset (Zhai et al. 2024c), which provides comprehensive scene-graph annotations for indoor scenes. This dataset includes 45K object instances and 15 types of relationships within bedrooms, dining rooms, and living rooms. Nodes in the scene graphs represent object categories, while edges indicate relationships between the nodes. In our experiments, we extracted corresponding images from the 3D-FUTURE dataset (Fu et al. 2021) based on node IDs to construct a Full-Modality Graph (node contains text and image). We then applied a random mask ratio to mask the text, images, and relationships between nodes in the Full-Modality Graph, producing the Mixed-Modality Graph.

Evaluation Metrics. We evaluate the scene-level and object-level fidelity of the synthesized 3D scenes. Scene-level fidelity is quantified using Fréchet Inception Distance (FID) (Heusel et al. 2017) and Kernel Inception Distance (KID) (Bińkowski et al. 2018), which measure the similarity between generated top-down renderings and real scenes renderings. For object-level fidelity, we assess the quality of generated object geometry using Minimum Matching Distance (MMD), Coverage (COV), and 1-Nearest Neighbor Accuracy (1-NNA) (Yang et al. 2019), all derived from Chamfer Distance (CD) (Fan, Su, and Guibas 2017).

Baselines. We compare our approach with three state-of-the-art scene synthesis methods: 1) **Graph-to-3D** (Dhamo et al. 2021), which generates 3D scenes directly from scene graphs using a GCN-based VAE; 2) **CommonScenes** (Zhai et al. 2024c), which converts scene graphs into controllable 3D scenes through a dual-branch framework with a VAE and LDM; 3) **EchoScene** (Zhai et al. 2024b), which employs a dual-branch diffusion model with an information echo mechanism for generating globally coherent 3D scenes from scene graphs.

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, FID_{CLIP}, and KID ($\times 0.001$), following the methodology in (Zhai et al. 2024c) (lower is better). **I** and **MM** represent nodes using image or mixed-modality respectively. **R** denotes the relationships of nodes. The best results are highlighted in **bold**.

Method	Shape Representation	Bedroom			Living room			Dining room		
		FID	FID_{CLIP}	KID	FID	FID_{CLIP}	KID	FID	FID_{CLIP}	KID
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
MMGDreamer (I+R)	echo2shape	**43.52**	**3.54**	**1.57**	**66.63**	**6.03**	**0.17**	**53.70**	**5.75**	**-0.09**

Table 2: **Object-level generation performance.** We present MMD ($\times 0.01$, $\downarrow$), COV($\%$, $\uparrow$), and 1-NNA($\%$, $\downarrow$) metrics to assess the quality and diversity of the generated shapes. **I** represents nodes using image representations. **R** denotes node relationships.

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 3: **Ablation studies.** Visual Enhancement Module and Relation Predictor are abbreviated as VEM and RP, respectively. The best results are highlighted in **bold**.

VEM	RP	FID	KID	mSG
		45.10	3.74	0.63
✓		43.50	3.27	0.63
	✓	44.07	3.43	0.83
✓	✓	**41.84**	**2.55**	**0.86**

Implementation Details. All experiments are performed on a single NVIDIA A100 GPU with 80 GB memory. We train our models using the AdamW optimizer, initializing the learning rate at 1×10^{-4} and utilizing a batch size of 128. The weighting factors for our loss components, α_1 and α_2, are consistently set to 1.0.

Scene Generation

Quantitative Comparison. We evaluate the realism of generated scenes using FID, FID_{CLIP}, and KID scores, as detailed in Tab. 1. MMGDreamer consistently outperforms the previous state-of-the-art method Echoscene, across all metrics when scene graph nodes are represented with either image or mixed-modality. Specifically, MMGDreamer (I+R) demonstrates a clear advantage over Echoscene. For instance, in living room generation, MMGDreamer (I+R) improves FID by 12%, FID_{CLIP} by 10%, and KID by 71%, highlighting its superior ability to control object geometry while enhancing overall scene realism. Furthermore, MMGDreamer (I+R) exhibits superior performance compared to MMGDreamer (MM+R), indicating that visual features offer substantial advantages over textual features in controlling and generating object geometry.

Qualitative Comparison. We present the generated results for different methods across various room types in Fig. 3. In comparison across different room types, our method MMGDreamer consistently demonstrates superior geometry control and visual fidelity in every scenario. For instance, in the bedroom, MMGDreamer accurately generates the bed and nightstands with higher geometric consistency, while other methods like Graph-to-3D and EchoScene display noticeable distortions and inconsistencies. In the dining room, both Graph-to-3D and EchoScene display significant deficiencies, particularly with chair backrests and the sideboard. In contrast, our method, MMGDreamer, not only preserves the correct geometry of these elements but also successfully generates the intricate details of objects placed on the sideboard. For the complex living room scene,

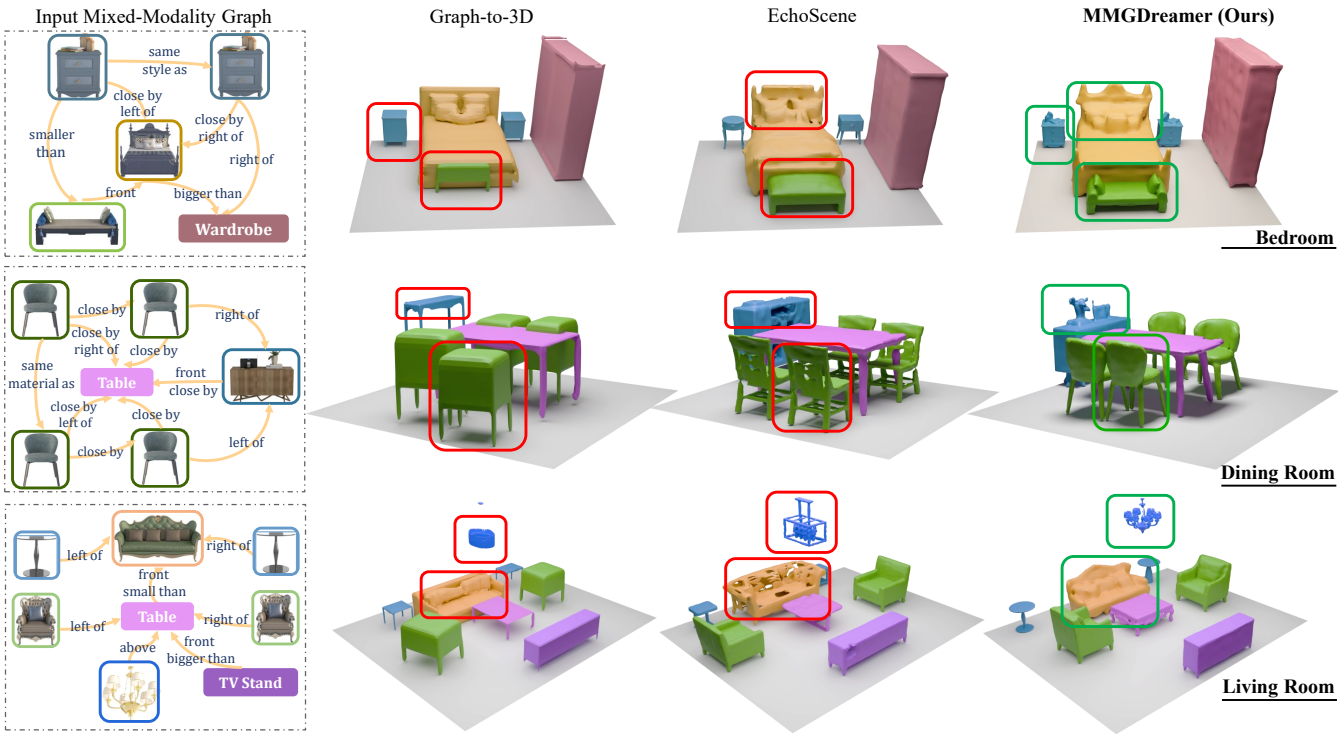

Figure 3: **Qualitative comparison** with other methods. The first column shows the input mixed-modality graph, which visualizes only the most critical edges in the scene. Red rectangles denote areas of inconsistency in the generated scenes, while green rectangles signify regions of consistent generation.

MMGDreamer accurately generates the sofa, coffee table, and lamp, maintaining a coherent spatial layout and ensuring a high degree of consistency between the generated objects and the input images. By contrast, other methods exhibit geometry errors in several pieces of furniture, such as the lamp and chair. Notably, the sofa generated by EchoScene contains numerous visible holes, significantly deviating from the actual object geometry.

Object Generation

We extend our analysis to the object level fidelity, following PointFlow (Yang et al. 2019), by reporting the MMD (×0.01), COV (%), and 1-nearest neighbor accuracy (1-NNA, %) metrics to assess per-object generation. As presented in Tab. 2, our method consistently surpasses the previous state-of-the-art across all object categories in both MMD and COV metrics. This result highlights MMGDreamer's geometric control capabilities, ensuring the precise generation of object geometry across various categories. The 1-NNA measures the distributional similarity between the generated objects and the ground truth, with values near 50% indicating better capturing of the shape distribution. Across most object categories, our method consistently outperforms EchoScene in terms of distributional similarity. Overall, MMGDreamer demonstrates superior geometric control, resulting in more consistent object-level generation compared to previous approaches.

Ablation Study

We utilize scene-level fidelity (FID and KID) and mean scene graph consistency (mSG) to quantitatively evaluate the effectiveness of different modules within MMGDreamer, as presented in Tab. 3. We observe that the configuration with VEM (second row) shows a significant decrease in FID and KID compared to the baseline (first row), indicating that VEM enhances the fidelity of scene generation. Additionally, when the RP module is introduced (third row), there is a notable improvement in mSG, demonstrating that RP effectively predicts relationships between objects, resulting in more coherent scene layouts. It is evident that including both VEM and RP achieves the best performance across all metrics, highlighting the complementary benefits of these modules in producing high-quality scene generation.

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. Extensive experiments demonstrate that MMGDreamer significantly outperforms existing methods, achieving state-of-the-art results in scene fidelity and object geometry controllability.

Acknowledgments

The authors would like to thank the anonymous reviewers for their comments. This work was supported by the National Key R&D Program of China under Grant 2023YFB2703800 and the Beijing Natural Science Foundation under Funding No. IS23055. The contact author is Zhen Xiao and Chao Zhang.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Basioti, K.; Abdelsalam, M. A.; Fancellu, F.; Pavlovic, V.; and Fazly, A. 2024. CIC-BART-SSA: Controllable Image Captioning with Structured Semantic Augmentation. *arXiv preprint arXiv:2407.11393*.

Bautista, M. A.; Guo, P.; Abnar, S.; Talbott, W.; Toshev, A.; Chen, Z.; Dinh, L.; Zhai, S.; Goh, H.; Ulbricht, D.; et al. 2022. Gaudi: A neural architect for immersive 3d scene generation. *Advances in Neural Information Processing Systems*, 35: 25102–25116.

Bińkowski, M.; Sutherland, D. J.; Arbel, M.; and Gretton, A. 2018. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.

Çelen, A.; Han, G.; Schindler, K.; Van Gool, L.; Armeni, I.; Obukhov, A.; and Wang, X. 2024. I-design: Personalized llm interior designer. *arXiv preprint arXiv:2404.02838*.

Cong, Y.; Yi, J.; Rosenhahn, B.; and Yang, M. Y. 2023. Ssgvs: Semantic scene graph-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2555–2565.

Curless, B.; and Levoy, M. 1996. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 303–312.

Dhamo, H.; Manhardt, F.; Navab, N.; and Tombari, F. 2021. Graph-to-3d: End-to-end generation and manipulation of 3d scenes using scene graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16352–16361.

Engelmann, F.; Rematas, K.; Leibe, B.; and Ferrari, V. 2021. From points to multi-object 3D reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4588–4597.

Epstein, D.; Poole, B.; Mildenhall, B.; Efros, A. A.; and Holynski, A. 2024. Disentangled 3d scene generation with layout learning. *arXiv preprint arXiv:2402.16936*.

Fan, H.; Su, H.; and Guibas, L. J. 2017. 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*, 605–613.

Fang, C.; Hu, X.; Luo, K.; and Tan, P. 2023. Ctrl-room: Controllable text-to-3d room meshes generation with layout constraints. *arXiv preprint arXiv:2310.03602*.

Fu, H.; Jia, R.; Gao, L.; Gong, M.; Zhao, B.; Maybank, S.; and Tao, D. 2021. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129: 3313–3337.

Hara, T.; and Harada, T. 2024. MaGRITTe: Manipulative and Generative 3D Realization from Image, Topview and Text. *arXiv preprint arXiv:2404.00345*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Höllein, L.; Cao, A.; Owens, A.; Johnson, J.; and Nießner, M. 2023. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7909–7920.

Hu, Z.; Iscen, A.; Jain, A.; Kipf, T.; Yue, Y.; Ross, D. A.; Schmid, C.; and Fathi, A. 2024. SceneCraft: An LLM Agent for Synthesizing 3D Scenes as Blender Code. In *Forty-first International Conference on Machine Learning*.

Jang, E.; Gu, S.; and Poole, B. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Johnson, J.; Gupta, A.; and Fei-Fei, L. 2018. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1219–1228.

Jyothi, A. A.; Durand, T.; He, J.; Sigal, L.; and Mori, G. 2019. Layoutvae: Stochastic scene layout generation from a label set. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9895–9904.

Koch, S.; Hermosilla, P.; Vaskevicius, N.; Colosi, M.; and Ropinski, T. 2024. Lang3DSG: Language-based contrastive pre-training for 3D Scene Graph prediction. In *2024 International Conference on 3D Vision (3DV)*, 1037–1047. IEEE.

Liao, R.; Erler, M.; Wang, H.; Zhai, G.; Zhang, G.; Ma, Y.; and Tresp, V. 2024a. VideoINSTA: Zero-shot Long Video Understanding via Informative Spatial-Temporal Reasoning with LLMs. *arXiv preprint arXiv:2409.20365*.

Liao, R.; Jia, X.; Li, Y.; Ma, Y.; and Tresp, V. 2024b. GenTKG: Generative Forecasting on Temporal Knowledge Graph with Large Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 4303–4317.

Lin, C.; and Mu, Y. 2024. Instructscene: Instruction-driven 3d indoor scene synthesis with semantic graph prior. *arXiv preprint arXiv:2402.04717*.

Para, W.; Guerrero, P.; Kelly, T.; Guibas, L. J.; and Wonka, P. 2021. Generative layout modeling using constraint graphs. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6690–6700.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ren, X.; Huang, J.; Zeng, X.; Museth, K.; Fidler, S.; and Williams, F. 2024. Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4209–4219.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Rosinol, A.; Gupta, A.; Abate, M.; Shi, J.; and Carlone, L. 2020. 3D dynamic scene graphs: Actionable spatial perception with places, objects, and humans. *arXiv preprint arXiv:2002.06289*.

Schult, J.; Tsai, S.; Höllein, L.; Wu, B.; Wang, J.; Ma, C.-Y.; Li, K.; Wang, X.; Wimbauer, F.; He, Z.; et al. 2024. Controlroom3d: Room generation using semantic proxy rooms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6201–6210.

Strader, J.; Hughes, N.; Chen, W.; Speranzon, A.; and Carlone, L. 2024. Indoor and outdoor 3d scene graph generation via language-enabled spatial ontologies. *IEEE Robotics and Automation Letters*.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Wald, J.; Dhamo, H.; Navab, N.; and Tombari, F. 2020. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3961–3970.

Wang, G.; Wang, P.; Chen, Z.; Wang, W.; Loy, C. C.; and Liu, Z. 2023. Perf: Panoramic neural radiance field from a single panorama. *arXiv preprint arXiv:2310.16831*.

Wang, X.; Yeshwanth, C.; and Nießner, M. 2021. Sceneformer: Indoor scene generation with transformers. In *2021 International Conference on 3D Vision (3DV)*, 106–115. IEEE.

Wu, Y.; Wei, P.; and Lin, L. 2023. Scene graph to image synthesis via knowledge consensus. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2856–2865.

Xu, Y.; Chai, M.; Shi, Z.; Peng, S.; Skorokhodov, I.; Siarohin, A.; Yang, C.; Shen, Y.; Lee, H.-Y.; Zhou, B.; 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.

Yan, H.; Li, Y.; Wu, Z.; Chen, S.; Sun, W.; Shang, T.; Liu, W.; Chen, T.; Dai, X.; Ma, C.; et al. 2024. Frankenstein: Generating Semantic-Compositional 3D Scenes in One Tri-Plane. *arXiv preprint arXiv:2403.16210*.

Yang, G.; Huang, X.; Hao, Z.; Liu, M.-Y.; Belongie, S.; and Hariharan, B. 2019. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4541–4550.

Yang, Y.; Jia, B.; Zhi, P.; and Huang, S. 2024. Physcene: Physically interactable 3d scene synthesis for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16262–16272.

Zhai, G.; Cai, X.; Huang, D.; Di, Y.; Manhardt, F.; Tombari, F.; Navab, N.; and Busam, B. 2024a. Sg-bot: Object rearrangement via coarse-to-fine robotic imagination on scene graphs. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 4303–4310. IEEE.

Zhai, G.; Örnek, E. P.; Chen, D. Z.; Liao, R.; Di, Y.; Navab, N.; Tombari, F.; and Busam, B. 2024b. EchoScene: Indoor Scene Generation via Information Echo over Scene Graph Diffusion. *arXiv preprint arXiv:2405.00915*.

Zhai, G.; Örnek, E. P.; Wu, S.-C.; Di, Y.; Tombari, F.; Navab, N.; and Busam, B. 2024c. Commonscenes: Generating commonsense 3d indoor scenes with scene graphs. *Advances in Neural Information Processing Systems*, 36.

Zhang, Y.; Huang, H.; Xiong, Z.; Shen, Z.; Lin, G.; Wang, H.; and Vun, N. 2024. Style-Consistent 3D Indoor Scene Synthesis with Decoupled Objects. *arXiv preprint arXiv:2401.13203*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Zhou, Y.; While, Z.; and Kalogerakis, E. 2019. Scenegraphnet: Neural message passing for 3d indoor scene augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7384–7392.