Autoregressive Diffusion Modeling for Compositional Text-to-3D Generation

Zhenggang Tang[1,2*], **Yuehao Wang**[1,3*], **Yuchen Fan**[1], **Jun-Kun Chen**[1,2], **Yu-Ying Yeh**[1], **Kihyuk Sohn**[1], **Zhangyang Wang**[3], **Qixing Huang**[3], **Rakesh Ranjan**[1], **Dilin Wang**[1,†], **Zhicheng Yan**[1,†]

[1]Meta Reality Labs, [2]University of Illinois Urbana-Champaign, [3]University of Texas at Austin
[*]equal contribution, [†]project lead

Recent text-to-scene generation approaches largely reduced the manual efforts required to create 3D scenes. However, their focus is either to generate a scene layout or to generate objects, and few generate both. The generated scene layout is often simple even with LLM's help. Moreover, the generated scene is often inconsistent with the text input that contains non-trivial descriptions of the shape, appearance, and spatial arrangement of the objects. We present a new paradigm of sequential text-to-scene generation and propose a novel generative model for interactive scene creation. At the core is a 3D Autoregressive Diffusion model 3D-ARD+, which unifies the autoregressive generation over a multimodal token sequence and diffusion generation of next-object 3D latents. To generate the next object, the model uses one autoregressive step to generate the coarse-grained 3D latents in the scene space, conditioned on both the current seen text instructions and already synthesized 3D scene. It then uses a second step to generate the 3D latents in the smaller object space, which can be decoded into fine-grained object geometry and appearance. We curate a large dataset of 230K indoor scenes with paired text instructions for training. We evaluate 7B 3D-ARD+ on 50 challenging scenes, and showcase the model can generate and place objects following non-trivial spatial layout and semantics prescribed by the text instructions. Code will be released.

Correspondence: zyan3@meta.com
Website: https://xxx.github.io/

Meta

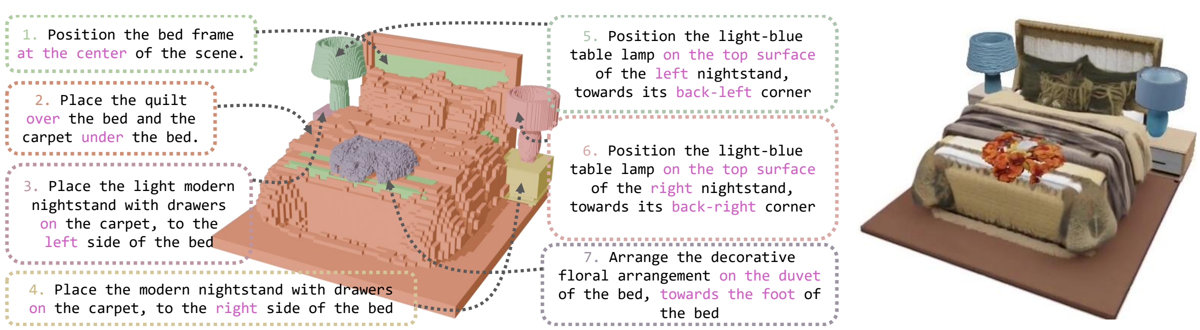

Figure 1 We present a 3D Autoregressive Diffusion model **3D-ARD+** to sequentially generate 3D objects from detailed text instructions, which not only describe the object shape and appearance, but also prescribe complex spatial relations between objects. 3D-ARD+ model generates a bedroom scene (**left**: occupancy voxel, **right**: appearance) precisely following the text instructions.

1 Introduction

The creation of immersive 3D scenes is crucial in gaming (UKCMA., 2022; Bhat et al., 2025; Hu et al., 2024; Li et al., 2025), virtual reality (Siddiqui et al., 2024; Zhou et al., 2024a, 2025c), and simulation for embodied AI (Yang et al., 2024c,b; Nasiriany et al., 2024; Deitke et al., 2022; Szot et al., 2021). This scene creation process is often interactive, where the user can compose a scene by sequentially adding objects with custom geometry, appearance, and spatial arrangement. Conventional workflows (Amirkhanov et al., 2025; Schönberger et al., 2016) often involve a time-consuming process that requires 3D artists to manually compose the scene, create detailed object geometry, and set up texture mapping. To reduce manual efforts, various text-to-scene approaches are proposed to synthesize 3D scenes from text input (Yang et al., 2024c; Li et al., 2024; Zhou et al., 2024b; Yang et al., 2024a; Bokhovkin et al., 2025; Ling et al., 2025; Zhou et al., 2025b; Wang et al., 2025b; Yang et al., 2025; Zhu et al., 2025), including models for layout generation (Tang et al., 2024), and object generation conditioned on layout (Zhang et al., 2024; Hu et al., 2024; Yan et al., 2024; Wu et al., 2024). However, fewer methods generate both (Vilesov et al., 2023; Fang et al., 2025).

For scene layout generation, earlier methods generate scene layout natively, but are limited to the layout of large objects only (*e.g.* sofa) and simple spatial relations, such as 2D layout (Fang et al., 2025) and basic relations (*e.g. a chair next to the table*) (Tang et al., 2024; Vilesov et al., 2023). They often do not handle more complex spatial relations, such as *Position the table lamp on the top surface of the right nightstand, towards its back-right corner* (Figure 1). More recent methods (Zhang et al., 2024; Wang et al., 2025b; Zhou et al., 2025b; Feng et al., 2023; Fu et al., 2024; Yang et al., 2024c; Li et al., 2024; Hong et al., 2025) exploit LLM (Hurst et al., 2024; Ouyang et al., 2022) to extract scene information from text input and generate a rough layout, which, however, often deviates from the text description, does not satisfy spatial relations, and needs further heuristic optimization based on spatial constraints and object interactions (*e.g.* Scene motif in (Pun et al., 2025), refinement in AnyHome (Fu et al., 2024)).

To generate objects conditioned on the scene layout, simple methods retrieve 3D assets from external sources (Tang et al., 2024; Yang et al., 2024b; Ling et al., 2025), which, however, often leads to the final scene inconsistent with the textual description. Advanced methods distill 3D representation from multi-view 2D images (Yang et al., 2024a; Zhang et al., 2024; Vilesov et al., 2023) or decode 3D representation from 3D latent codes generated by diffusion models (Wu et al., 2024; Yan et al., 2024). However, generated objects are often limited to a predefined set of object categories (Fang et al., 2025; Bokhovkin et al., 2025; Yan et al., 2024) or lack geometric details compared to text input (Wu et al., 2024).

To support the interactive scene creation with detailed object shape and appearance, we propose a novel *3D AutoRegressive Diffusion model (3D-ARD+)* to natively generate objects with different sizes and non-trivial spatial arrangement according to the sequential text input. When the text input prescribes fine details about the shape, appearance, and spatial relations of objects, we show that it is challenging for existing approaches, while our 3D-ARD+ model performs significantly better. The 3D-ARD+ model processes the text instructions sequentially. Each text instruction describes the shape, appearance and placement of a new object, and our 3D-ARD+ model autoregressively generates the placement, fine-grained geometry, and appearance of the new object.

Under the hood, for each text instruction, our model processes the text tokens encoded from all seen text input, and the 3D latents (Xiang et al., 2025) of the current scene to predict the 3D latents of the next object in the large scene. A subsequent generation step is used to generate the 3D latents of the new object in the smaller object space, which can be decoded into fine-grained 3D Gaussians. The newly generated object is tokenized and appended to the multimodal token sequence to condition the future object generation in an autoregressive manner. The 3D-ARD+ model adopts the DiT transformer architecture (Peebles and Xie, 2023) with causal attention between text- and 3D tokens, and unrestricted attention between 3D tokens.

To train the 3D-ARD+ model, we curated a proprietary dataset consisting of 230K indoor scenes. We prompt a public VLM (Comanici et al., 2025) to generate step-by-step text instructions to mimic the scene creation process. On the evaluation set, which contains 50 sets of text instructions for composing non-trivial multi-object scenes, we extensively compare our 3D-ARD+ model with several competing methods (Huang et al., 2025b; Xiang et al., 2025), and validate the 3D-ARD+ model performs significantly better in preserving the spatial relations and generating object geometry and appearance, even when object scale varies largely.

We summarize our contributions as follows:

- We present a new paradigm of sequential text-to-scene generation and curate an evaluation set of compositional text instructions captioned for 50 indoor scenes.
- We propose a novel 3D AutoRegressive Diffusion model to autoregressively generate the shape, texture, and placement of the next object conditioned on the already synthesized scene and all seen text instructions.

- We develop a data pipeline to collect a large dataset of 230K indoor scenes with paired text-scene data.
- On our challenging evaluation set, we demonstrate that the 3D-ARD+ model outperforms other methods by a large margin in composing multi-object scenes.

2 Related Work

2.1 Text-to-3D Generation

Early Text-to-3D approaches generate 3D objects by distilling 2D diffusion priors (Rombach et al., 2022; Peebles and Xie, 2023), without any 3D training data, including DreamFusion (Poole et al., 2023), LucidDreamer (Wang et al., 2023b), and many others (Lin et al., 2023; Wang et al., 2023a; Zhu et al., 2024; Liang et al., 2024). Multi-view diffusion models (Shi et al., 2024b; Liu et al., 2024; Long et al., 2024) directly generate pose-consistent image views, which can be used to reconstruct 3D objects. Recent 3D generative models (Nichol et al., 2022; Jun and Nichol, 2023; Xiang et al., 2025) learn to directly map text to 3D latents or explicit 3D representations but are not capable of distinguishing individual objects and preserving the spatial arrangement prescribed in the text. It is still challenging to apply such Text-to-3D approaches to the task of interactive multi-object scene generation, where the user still needs to manually place the generated individual objects into the 3D scene.

2.2 3D Indoor Scene Generation

Indoor scene generation approaches often leverage LLMs (Ouyang et al., 2022; Hurst et al., 2024), 2D (Podell et al., 2023; Rombach et al., 2022), and 3D generative models (Poole et al., 2023; Xiang et al., 2025). While some are focused on scene layout generation, others focused on generating objects or both.

For scene layout generation, autoregressive (Paschalidou et al., 2021) and diffusion (Tang et al., 2024; Fang et al., 2025) models are often used to generate a scene code and 3D object attributes. CG3D (Vilesov et al., 2023) uses a diffusion model for compositional 3D object generation, but is often limited to modeling simple spatial relations. With the tremendous advances in LLMs, many work exploits them for scene layout generation (Feng et al., 2023; Zhang et al., 2024; Wang et al., 2025b; Zhou et al., 2025b; Yang et al., 2024c; Li et al., 2024; Hong et al., 2025; Çelen et al., 2024; Ran et al., 2025). Although LLMs improve the scene diversity, the result is often not well aligned with the text input and requires further optimizations (Yang et al., 2024c; Çelen et al., 2024; Ran et al., 2025). AnyHome (Fu et al., 2024) prompts the LLM to convert the input text into structured representations, but still rectifies the room layout by Score Distillation Sampling (Poole et al., 2023). HSM (Pun et al., 2025) uses a VLM to extract room type and objects from the text, but still needs multiple steps to generate the final layout, such as extracting support region, generating scene motif and optimizing the room layout. Our 3D-ARD+ model natively generates the next object conditioned on the seen text instructions and already generated objects, and thus places the next object in the scene consistent with the input.

To create 3D objects in the scene, (Tang et al., 2024; Yang et al., 2024b; Ling et al., 2025) retrieve 3D assets from asset libraries (Deitke et al., 2023b,a), but it lacks coherence to text instruction or between objects. Instead, recent works generate 3D objects (Yang et al., 2024a; Zhang et al., 2024; Vilesov et al., 2023; Wu et al., 2024; Yan et al., 2024; Fang et al., 2025; Bokhovkin et al., 2025). BlockFusion (Wu et al., 2024) and SceneFactor (Bokhovkin et al., 2025) generate 3D objects holistically as a scene, thus lacking geometric details for individual objects, and requiring further refinement. SceneCraft (Yang et al., 2024a) reconstructs 3D NeRF representation from multi-view images. CG3D (Vilesov et al., 2023), DreamScene (Li et al., 2024) generate 3D objects with variants of score distillation sampling (Poole et al., 2023). In contrast, our 3D-ARD+ generates 3D latents, which can be decoded into 3D representation (*e.g.*, 3D Gaussian) of individual objects in the scene step-by-step, thereby maintaining the fine-grained details and high fidelity.

With the rise of video diffusion models (VDMs), generating 3D scenes from images or videos has been studied. ArtiScene (Gu et al., 2025) extracts 3D attributes from isometric scene layouts and generates 3D objects via image-to-3D model. StarGen (Zhai et al., 2025), HunyuanWorld-Voyager (Huang et al., 2025a) propose long-range, pose-controllable VDMs whose frames can be turned into 3D Gaussian splats. However, these

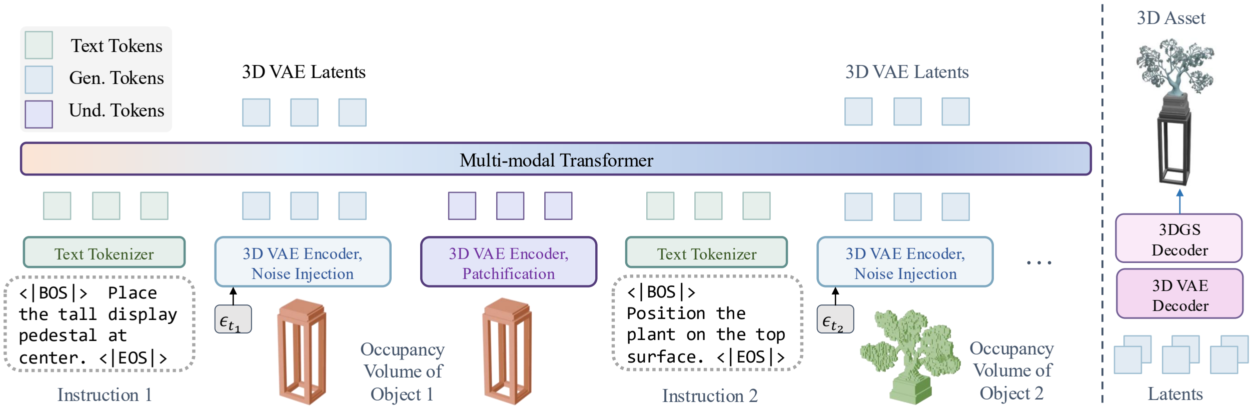

Figure 2 Overview of 3D-ARD model for coarse-grained scene generation. Left: at training time, the model takes text tokens, 3D understanding tokens, and noised 3D generation tokens as input, and predicts a time-dependent noise. **Right**: at inference time, the model iteratively transforms a random noise into 3D latents, which can be decoded by 3D VAE decoder and 3D Gaussian (3DGS) decoder to generate a 3D object.

two-stage approaches are susceptible to error accumulation, and the quality of 3D reconstruction is limited by the fidelity of the intermediate images or videos.

2.3 Multimodal Generative Models

Since the success of visual generation models based on the text prompt (Ramesh et al., 2021; Saharia et al., 2022; Rombach et al., 2022; Dai et al., 2023), it has been studied to generate visual contents beyond the text input, such as reference images. Earlier work focused on building a specialized model for each set of reference images (Gal et al., 2023; Ruiz et al., 2023). However, such approachs are expensive and limited to making reference to only a few concepts.

Following the success of LLMs, multimodal generative models are introduced (Hurst et al., 2024; Comanici et al., 2025; Zhou et al., 2025a; Team, 2024; Shi et al., 2024a; Chen et al., 2025; Xie et al., 2025; Pan et al., 2025; Wang et al., 2025a; Wu et al., 2025a; Deng et al., 2025). Compared to earlier works based on fine-tuning, multimodal generative models are fast and flexible, as it takes multimodal inputs (*e.g.* reference image) in an in-context manner to generate multimodal outputs. Chameleon (Team, 2024), ILLUME (Wang et al., 2025a), MUSE-VL (Xie et al., 2025), and Janus-family (Wu et al., 2025a; Chen et al., 2025) employ a discrete image tokenizer so that both text and vision modalities can be modeled using a single autoregressive transformer. On the other hand, Transfusion (Zhou et al., 2025a), LMfusion (Shi et al., 2024a), MetaQuery (Pan et al., 2025) and Bagel (Deng et al., 2025) develop models by combining the best of both worlds, where the discrete (*e.g.* text) token is processed using next-token prediction, while the continous (*e.g.* image) token is processed via diffusion. The proposed 3D-ARD+ is inspired by latter that combines next-token prediction and diffusion for multimodal generation, but is designed to generate 3D representation directly from our autoregressive-diffusion model.

3 Method

We tackle the task of sequential text-to-scene generation, and the goal is to generate individual objects in a multi-step process based on sequential text instructions, which often prescribes the shape, appearance, functional use and spatial arrangement of the new object at each step. We do not generate walls, floors and ceilings since they can be easily generated by prior methods (Raistrick et al., 2024; Yang et al., 2024c; Fu et al., 2024; Pun et al., 2025).

We propose a novel 3D Autoregressive Diffusion model **3D-ARD** for coarse-grained scene generation (Figure 2), and further extend it into a **3D-ARD+** model for fine-grained scene generation (Figure 5). Both are trained using a diffusion objective (Liu et al., 2023). We address three key challenges below. **1)**: *How to build a model to generate the next object with coarse shape and appearance conditioned on the text input and the*

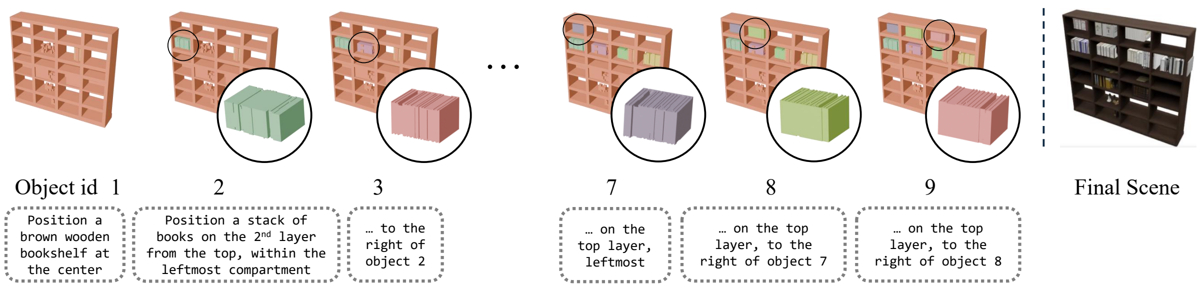

Figure 3 An example of step-by-step generation. Our 3D-ARD+ model closely follows the text instructions and add a total of 8 stacks of books to different compartments in a bookshelf.

already synthesized 3D scene (if any)? In Section 3.1, we present the 3D-ARD model architecture. **2)**: *How to generate objects with fine geometric details and appearance when the object size varies and the scene is often much larger?* In Section 3.2, we present an extended 3D-ARD+ model for refining the object geometry using extra refinement steps. **3)** *The lack of training data with paired scene and sequential text input.* In Section 3.3, we introduce our data pipeline to curate a large-scale indoor scene dataset with paired text instructions.

3.1 3D-ARD: Autoregressive 3D Diffusion

We denote the sequential text instructions as $\{T_t\}_{t=1}^{N}$ where N is the total steps. The text instruction T_t for generating the next object often not only describes the shape and appearance of the object, but also its spatial arrangement. An example from Figure 1 is *"Position the light-blue table lamp on the top surface of the right nightstand, towards its back-right corner."* Therefore, the next-object generation should be conditioned on all seen text instructions and the already synthetized 3D objects (if any). Inspired by Transfusion (Zhou et al., 2025a), we build a multi-modal transformer model that simultaneously processes text tokens X_t^T, 3D understanding tokens X_t^U and 3D generation tokens X_t^G at each step t.

Text tokens. We tokenize the text string T_t into a sequence of discrete tokens with the BAGEL tokenizer (Deng et al., 2025), and use standard embedding layers to convert tokens into vectors $\text{Embed}(T_t)$ of dimension C. At each step, we concatenate text tokens of all seen text instructions to obtain X_t^T.

3D understanding tokens. At training time, for each of the existing objects $\{O_{t'}\}_{t'=1}^{t-1}$ in the step t, we take a 3D binary volume $V_{t'} \in \{0,1\}^{M\times M\times M}$, which represents the occupancy of object $O_{t'}$ in the whole scene, and use a VAE encoder to encode it into low-resolution 3D latents $S_{t'} \in \mathbb{R}^{D\times D\times D\times C_S}$. To reduce the number of tokens, we process $S_{t'}$ using a patchification layer with non-overlapping patches of size 2, followed by a linear projection layer to align with the text embedding dimension C. The 3D understanding tokens X_t^U include tokens of all existing objects.

3D generation tokens. Inspired by TRELLIS (Xiang et al., 2025), the occupany of an object in the scene can be represented by a list of active voxels $\{p_i\}_{i=1}^{L}$, where p_i is the position index of a voxel, and L the total number of active voxels. The sparse voxels $\{p_i\}_{i=1}^{L}$ are converted into a dense binary 3D volume $V \in \{0,1\}^{M\times M\times M}$, which is further encoded by a 3D VAE encoder into low-resolution 3D latents $S \in \mathbb{R}^{D\times D\times D\times C_S}$. During model training, we obtain the 3D generation tokens in the current step by linearly projecting a noised version of S into X_t^G. Note the generation tokens will be later denoised and decoded to predict the occupancy of next object in the scene.

3.1.1 Training Recipe

We train the 3D-ARD model on a curated training dataset (Section 3.3) with paired scene and sequential text instructions. Over the generation steps $\{t\}$, we autoregressively predict the denoised 3D latents of individual objects, which can be decoded into a 3D occupancy volume V in the scene. Unlike Transfusion (Zhou et al., 2025a), which predicts both text token and image patches and thus applies objectives on all output tokens, our goal is to generate the next object in the scene. Therefore, we apply diffusion objective to the denoised 3D generation tokens only. Specifically, we model the 3D latents distribution using the Rectified flow model (Liu et al., 2023), where in the forward pass a noised sample is obtained based on a time-dependent linear interpolation $\boldsymbol{x}(s) = (1-s)\boldsymbol{x} + s\boldsymbol{\epsilon}$ between a sample $\boldsymbol{x}$ and a random noise $\boldsymbol{\epsilon}$. In the backward pass, the noised sample is denoised according to a time-dependent flow $\boldsymbol{v}(\boldsymbol{x}, s) = \nabla_s(x)$, which is approximated by the transformer backbone trained with the conditional flow matching objective below.

$$\mathcal{L}(\theta) = \mathbb{E}_{s,\boldsymbol{x}_0,\boldsymbol{\epsilon}} \|\boldsymbol{v}_\theta(\boldsymbol{x}, s) - (\boldsymbol{\epsilon} - \boldsymbol{x}_0)\|_2^2 \tag{1}$$

where θ denotes the learnable parameters.

3.1.2 Transformer Backbone

The majority of the 3D-ARD model's parameters θ are with the multimodal transformer backbone, which consists of 28 blocks with self-attention (Peebles and Xie, 2023). To explicitly reveal different types of tokens in the sequence, we insert BOS and EOS tokens to denote the beginning and end of text tokens. Similarly, we insert BO3D and EO3D tokens for 3D understanding and generation tokens.

3D-ARD Transformer attention. We implement a generalized causal attention between three types of input tokens (Figure 4). For text tokens, it uses standard causal attention, including attention to the text tokens and 3D understanding tokens from earlier steps.

For 3D understanding tokens, it uses the standard causal attention to text tokens, but unrestricted attention to themselves, allowing every understanding token to attend to every other understanding token in the scene. For 3D generation tokens, it uses a standard causal attention to text and 3D understanding tokens, since the generation of next object should be conditioned on all seen text and already synthesized objects. They have unrestricted attention to themselves, allowing every generation token to attend to every other generation token.

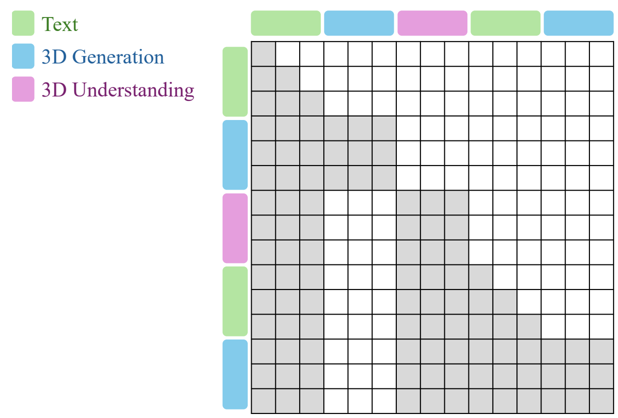

Figure 4 The generalized causal attention used by 3D-ARD model. Tokens along the horizontal and vertical directions are input and output tokens, respectively.

3.1.3 Model Inference

At inference time, for each text instruction T_t, we first sample a random noise $\boldsymbol{\epsilon}$ as $\boldsymbol{x}(1)$, and iteratively follow the approximate flow $\boldsymbol{v}_\theta(\boldsymbol{x}, s)$, conditioned on all past text tokens and 3D understanding tokens, to update the sample $\boldsymbol{x}(s)$ until we obtain a denoised sample $\boldsymbol{x}(0)$. We decode $\boldsymbol{x}(0)$ into a binary volume V via a 3D VAE decoder to denote the occupancy of the object in the scene. To obtain the geometry and appearance of the final object, we extract the active voxels $\{p_i\}_{i=1}^{L}$ from V, iteratively denoise a randomly initialized noise by running the off-the-shelve TRELLIS 3D VAE decoder based on sparse flow transformer (Xiang et al., 2025) conditioned on the current object text description to obtain the structured latent $z = \{(z_i, p_i)\}_{i=1}^{L}$, which can be decoded into 3D Gaussians using the TRELLIS 3DGS decoder (See Figure 1).

3.2 3D-ARD+: Fine-grained Scene Generation

Due to computational constraints, the resolution of the generated object in the scene space $V \in \{0,1\}^{M\times M\times M}$ is limited (M=64 in our experiments), resulting in only coarse-grained object geometry for common objects (Figure 7 Top). To address this limitation, we extend the 3D-ARD model by adding an extra step to generate 3D latents of the same resolution $D\times D\times D$ in the smaller object space after each current generation step, and the resulting model is referred to as **3D-ARD+** model. It uses a generation process of $2N$ steps for sequential

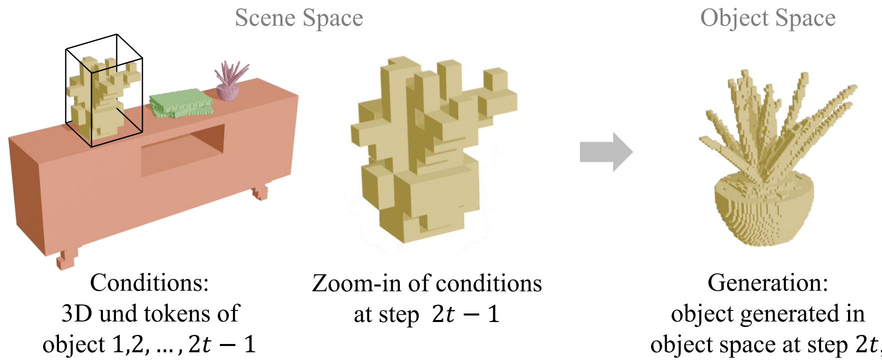

Figure 5 Fine-grained object generation in 3D-ARD+ model. 3D-ARD+ model generates 3D latents in the object space to obtain more fine-grained geometry.

text instructions $\{T_t\}_{t=1}^{N}$ (Figure 5). Below, we present details on how to prepare tokens at even steps $\{2t\}$, where fine-grained objects are generated.

Text tokens. At an even step $2t$, we take text tokens of the instruction T_t already prepared at the step $2t-1$, as well as two new special tokens BOR and EOR, which denote the beginning and end of refinement text, as the text tokens at the step $2t$.

3D understanding tokens. At an even step $2t$, we first prepare 3D understanding tokens from all past odd steps $\{2t'-1\}_{t'=1}^{t}$ as in section 3.1. For all past even steps $\{2t'\}_{t'=1}^{t-1}$, we take the groundtruth dense binary 3D volume $V_{2t'}$ in the object space, and use a 3D VAE encoder to encode it into a low-resolution 3D latents $S_{2t'}$. Similar to section 3.1, latents $S_{2t'}$ are further patchified and linearly projected to reduce the token numbers and align the feature dimension.

3D generation tokens. To prepare 3D generation tokens, we take the groundtruth binary occupancy volume V in the object space, encode it into low-resolution 3D latents S via a 3D VAE encoder, and apply the Rectified Flow time-dependent interpolation to obtain a noisy sample. As in Section 3.1.1, we further linearly project it to obtain 3D generation tokens.

Model training. 3D-ARD+ model at both odd and even generation steps also attaches the conditional flow matching objective (Equation 1) to the denoised 3D generation tokens to train the transformer backbone.

Model inference. Similarly to the inference of the 3D-ARD model in Section 3.1.3, the 3D latents generated in the local object space are decoded into a binary occupancy volume V, where the active voxels with normalized coordinates are extracted to be used by the TRELLIS 3D VAE based on sparse flow transformer to generate structured latents $z = \{(z_i, p_i)\}_{i=1}^{L}$. To put the fine occupancy volume in the scene space, we first calculate the bounding box of the coarse one and then transform the fine one accordingly.

3.3 Dataset Construction

To enable sequential scene generation, we construct a high-quality indoor dataset with step-wise assembly instructions. The raw data is from a proprietary indoor dataset, which consists of parts of a scene but lacks of assembling instructions. To generate the instructions, our data pipeline consists of four main stages (Figure 6).

#1 View selection. Given multi-view renderings of each object group, we leverage a public VLM API (Comanici et al., 2025) to identify the canonical front view that maximizes visual informativeness while minimizing occlusion. The model analyzes specific cues such as furniture orientation (*e.g.* cabinet doors, bed headboards) to determine the optimal viewing angle for subsequent labeling.

#2 Object captioning. For each object in the scene, we generate multi-granularity textual descriptions by presenting the VLM with both the complete scene view and individual object-focused views. The captions capture essential attributes including object type, material properties, color, geometric characteristics, and critically structural elements that support other objects.

#3 Spatial relationship analysis. We incorporate 3D bounding box data (including dimensions, center positions, and spatial extents) to provide a precise geometric context for this spatial reasoning step. A top-down visualization is also generated to facilitate it.

#4 Assembly plan generation. Combining the visual observations, object captions, geometric context, and spatial reasoning, the VLM generates a sequential assembly plan consisting of step-by-step instructions that specifies the placement order from foundational objects to dependent components. The procedures help VLM to generate accurate assembly instructions.

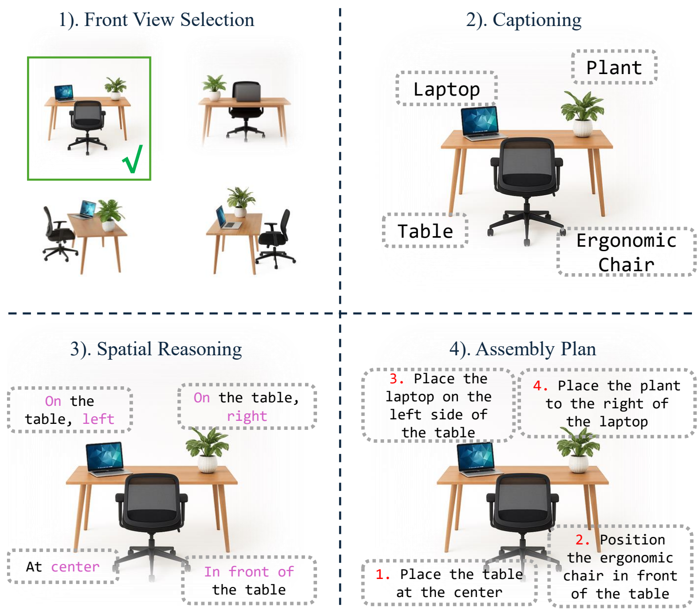

Figure 6 Dataset construction pipeline includes four steps: (1) Front view selection, (2) Captioning, (3) Spatial reasoning, and (4) Generating assembly plan.

4 Experiments

4.1 Implementation Details

We implement our approach in PyTorch (Paszke et al., 2019). We use the dense binary occupancy volume $V \in \{0,1\}^{M\times M\times M}$ with resolution M=64 to represent the occurrence of objects in the scene space or the smaller object space. An off-the-shelve 3D VAE from TRELLIS (Xiang et al., 2025) is used to encode V into a low-resolution continuous volume $S \in \mathbb{R}^{D\times D\times D\times C_S}$, where $D = 16$ and $C_S = 8$. All text tokens, 3D understanding tokens, and 3D generation tokens use the feature dimension 128.

3D-ARD+ model training. We use pre-trained multimodal transformer model from Bagel (Deng et al., 2025), which contains 7B active parameters and 28 self-attention building blocks. We finetune it on our curated indoor data for 120K steps with learning rate $1e-4$ using 128 Nvidia H100 GPUs for a week. The maximum token size per sequence is 20,480.

3D-ARD+ model inference. KV pairs of text tokens for all seen text instructions are stored in the KV cache (Pope et al., 2022). KV pairs of denoised 3D generation tokens are also stored. We utilize the Euler sampler for generation, employing 50 sampling steps. The CFG coefficient (Ho and Salimans, 2022) is set to 4 for text conditions and 2 for 3D condition tokens.

Indoor training data. Our dataset consists of 230K indoor scenes from typical room types (bedrooms, living rooms, kitchens, dining areas, bathrooms). Each scene contains 2–15 parts with diverse compositions: adjacent furniture (*e.g.* bed with nightstands), furniture-object assemblies (*e.g.* table with tableware), or grouped small objects. Each scene is normalized together to be within the unit bounding box. Each part has a corresponding assembly instruction (several to 40 words) describing its spatial placement.

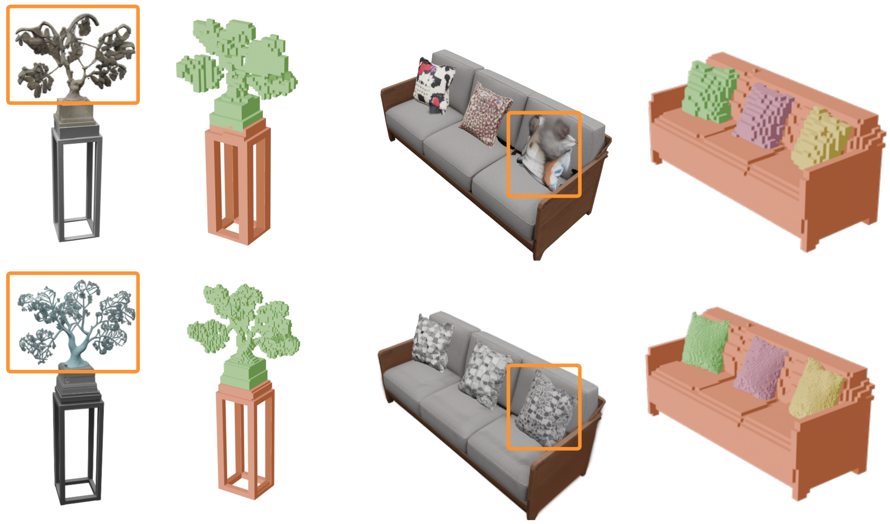

Figure 7 Comparing results from 3D-ARD and 3D-ARD+ models. Top: coarse-grained; Bottom: fine-grained.

4.2 Evaluation Settings

Our evaluation set consists of sequential text instructions captioned by a public VLM API (Comanici et al., 2025), as detailed in Section 3.3. See examples in the appendix.

4.2.1 Evaluation dataset

We select 50 unique and challenging scenes from the curated dataset for evaluation. These scenes span diverse room types (*e.g.* kitchen, bedroom, living room), varying spatial scales (from groups of large furnitures to groups of small objects), and different object counts, ensuring comprehensive coverage of compositional complexity and spatial reasoning challenges.

4.2.2 Baselines

We consider the following four baselines.

TRELLIS-text XL. is the largest Text-to-3D model TRELLIS (Xiang et al., 2025) open-sourced with 2B parameters. To prepare the text input, we use a public VML API (Comanici et al., 2025) to summarize all steps into one instruction.

MIDI: Multi-Instance Diffusion. is a recent image-to-scene model to generate multi-object scenes by segmenting the scene image using Grounded-SAM (Ren and et al., 2024), and applying a multi-instance diffusion model (Huang et al., 2025b). To prepare the image input, we summarize all text instructions into an overall description of the scene, and use a public text-to-image API (Comanici et al., 2025) to generate a scene image.

BB.+TRELLIS. We reuse the 3D bounding boxes in the test set. Then we use the TRELLIS-text XL model to generate 3D objects from object captions, and fit the generated object to the bounding box afterwards.

LLM-layout+TRELLIS. In this baseline, we first use the text instructions to prompt a public LLM API for generating detailed object captions and predicting 3D bounding boxes. Then we use the TRELLIS-text XL model to generate actual 3D objects from those captions, rescale and position each mesh to fit its bounding box.

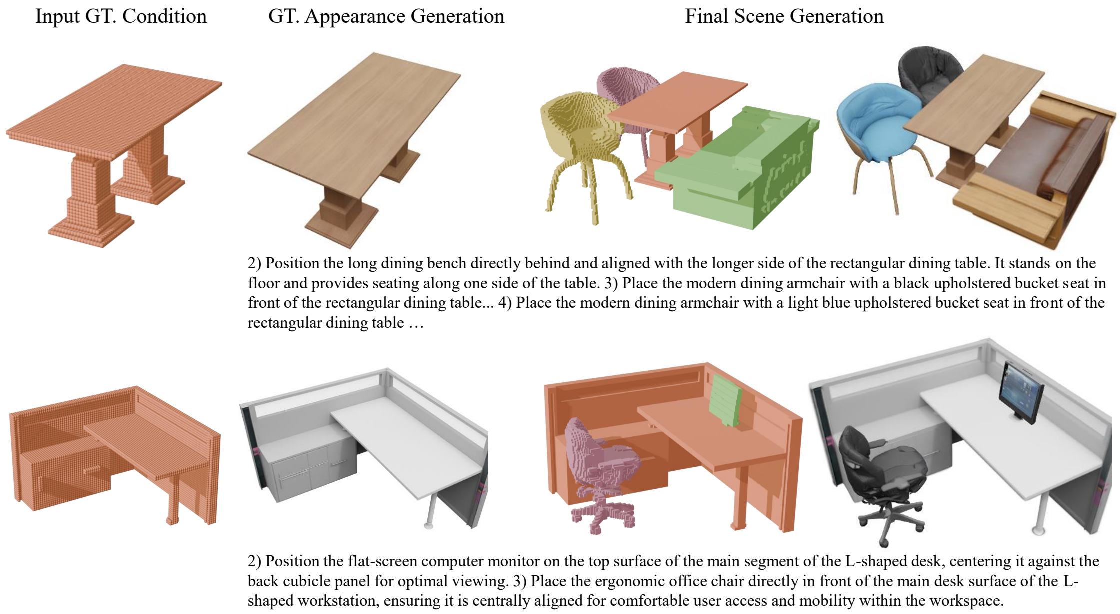

Figure 8 Conditioned next-object generation. Given the ground-truth foundamental object (table or desk) and subsequent textual instructions, our model generates subsequent objects properly.

4.3 Qualitative Results

3D-ARD+ sequential generation results. As shown in Figure 3, 3D-ARD+ closely follows text instructions at each step, and can generate object geometry and appearance close to the text instructions while also placing the object based on the prescribed spatial arrangement.

Comparing 3D-ARD and 3D-ARD+ models. In Figure 7, we show 3D-ARD+ generates more fine-grained objects with the designed extra refinement step in the left example. The top right example shows that generating appearance directly on coarse geometry will be likely to cause artifacts in the resulting texture.

Comparisons with baselines. In Figure 10, we qualitatively compare our 3D-ARD+ model with four baselines as our primiary evaluation. *TRELLIS-text XL* will hallucinate unrelated objects (*e.g.* paper towel in row #1) or generate objects inconsistent with the text input (*e.g.* more than expected pillows in row #2). *MIDI* requires an image aligned with the text as the input. We empirically observe that when images generated from text instructions by a public VLM API (Comanici et al., 2025) are used as input, MIDI may still fail to produce plausible object shapes and placements. See more details in the supplementary. For *BB.+TRELLIS*, it is difficult to avoid object collision after placing individual objects, generated by TRELLIS-text XL, into the scene according to the groundtruth bounding box (*e.g.* row #3). For *LLM-layout+TRELLIS*, LLM can predict 3D bounding box of individual objects poorly for complex text input (*e.g.* row #1, #3). Even the subsequent generation of individual objects by TRELLIS-text XL are plausible, the spatial arrangement in the composed scene still significantly deviates from the text instructions. In contrast, our method successfully captures object types, shapes, appearance, and their spatial locations as specified in the text instructions, resulting in meaningful compositional 3D generation.

Conditioned Next-Object Generation Our model supports conditioned generation where the first object is fixed and subsequent objects are generated based on varying text prompts. Figure 8 demonstrates this capability.

Diversity of Generation To demonstrate the generative diversity of our model, we provide multiple generations with different random seeds from identical text instructions. Figure 9 shows that given the same prompt, our model produces varied yet semantically consistent results.

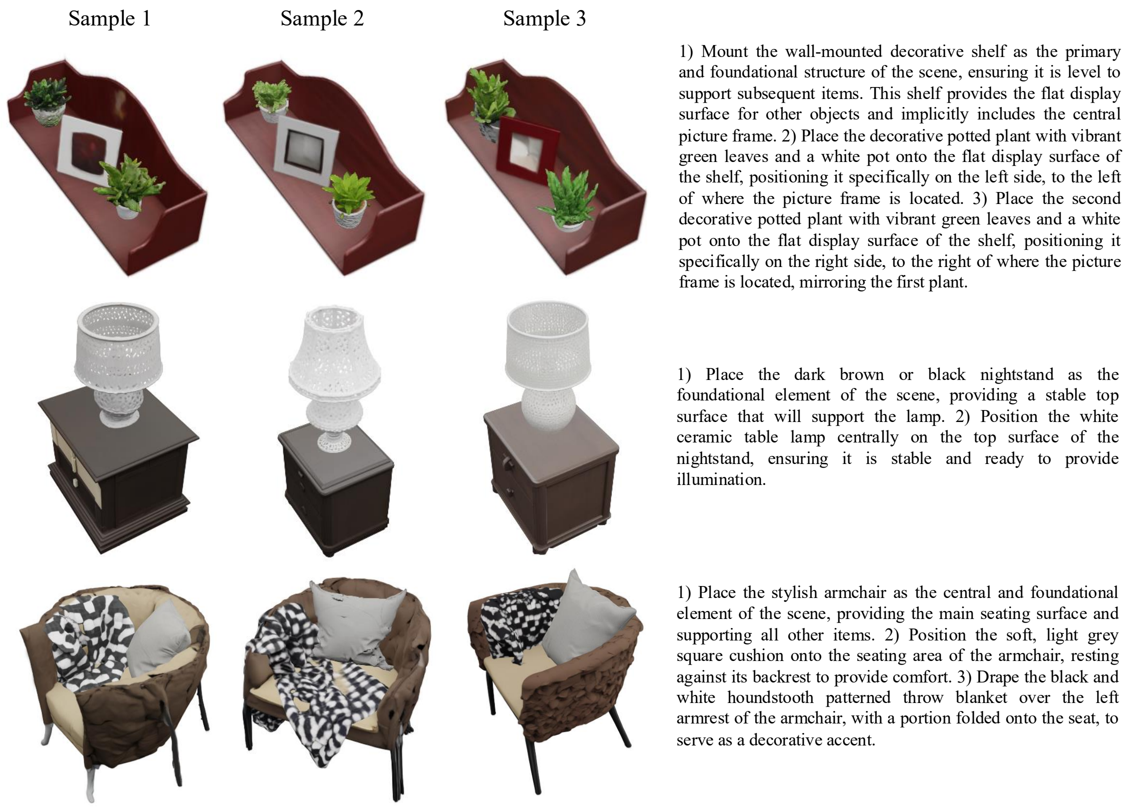

Figure 9 Diversity of generation results. Three samples are generated for each sequence of text instructions, illustrating the variations of output produced by our method. The first-row results are generated via conditioning on the ground-truth geometry of the base object. Results in the second and third rows are generated from scratch.

4.4 Quantitative Comparisons

Evaluation metrics. We conduct the evaluation of 3D generation on their rendered images. Following previous work (Xiang et al., 2025), we employ kernel distance metrics, including Kernel Inception Distance (KID) (Bińkowski et al., 2018) using InceptionV3, and Kernel Distance with DINOv2 (KDD) encoders. For each scene in our training set, we randomly render images from the front, left, right, and back views to construct the reference batch. Similarly, we render these four views for each baseline in the test set to form the generation batch. To measure the consistency between text descriptions and generated scenes, we use CLIP scores, comparing both text and image references. To obtain a single reference text description (rather than a list of instructions) and image for each scene, we first input the textual instructions into a public VLM (Comanici et al., 2025) to generate the corresponding images. We then query it to describe these generated images, using the resulting descriptions as reference text prompts.

Results. We present our quantitative results in Table 1 as a secondary evaluation. Our method consistently outperforms all baseline approaches across every metric, aligning with our visual comparison. These findings offer numerical validation of our model's effectiveness, demonstrating superior overall generation quality and enhanced consistency with text descriptions.

5 Limitations and Future Work

While our proposed method demonstrates promising results in language-conditioned 3D scene generation, we acknowledge two main limitations that present opportunities for future research. 1. Despite our explicit spatial condition, the generated layouts are not always perfect. Specifically, overlapping or gaps between two adjacent objects are not always resolved correctly. While our assembly plan provides sequential placement

Table 1 Quantitative comparisons on the indoor evaluation set. We report several metrics, including Kernel Distance (Bińkowski et al., 2018) with Inception-v3 (KID) and DINOv2 encoders, CLIP-text, and CLIP-image scores (Radford et al., 2021).

	KID (↓)	KDD (↓)	CLIP-text (↑)	CLIP-image (↑)
MIDI	1.308	98.721	22.607	74.896
TRELLIS-text XL	1.543	96.928	22.086	73.306
BB.+TRELLIS	1.694	101.287	20.726	66.555
LLM-layout+TRELLIS	1.376	88.425	22.645	71.896
Ours	**1.244**	**49.021**	**27.447**	**76.103**

instructions and relative spatial relationships, the spatial condition representation (VAE tokens) may lack sufficient granularity to precisely encode fine-grained spatial positioning. Instead of relying solely on VAE tokens, we can explore more fine-grained encoding by focuing only on surface area in a higher resolution, or use other more elaborated encoders Wu et al. (2025b) that shows better spatial understanding accuracy. 2. The size of the first placed object (foundational element) might be too large relative to the intended scene bounds, leaving insufficient space to accommodate subsequent objects according to the assembly plan. This can lead to spatial constraint violations. To alleviate this, we can design an adaptive scene generation framework where the available spatial bounds adjust dynamically based on the scale of all objects generated up to now. This would enable the model to allocate appropriate space before putting the next object.

6 Conclusion

We present a novel Autoregressive 3D Diffusion model 3D-ARD+ for compositional text-to-3D generation. A large-scale indoor scene dataset with paired text instructions is curated to train the 3D-ARD+ model. On a challenging evaluation set of 50 text instruction sets, the proposed 3D-ARD+ model significantly outperforms other competing methods in both visual inspection and quantitative evaluations.

Figure 10 Qualitative comparison. Visual results of our method and existing single-object generation and multi-object generation approaches are presented with corresponding input text instructions and reference images generated by a public VLM API (Comanici et al., 2025). **Bold text** highlights object kinds, *italic text* shape/appearance descriptions, and underline locations.

Figure 11 Additional qualitative comparison results. We present additional visual results comparing our method with baseline approaches. Grey boxes labeled "FAILED" indicate that the corresponding baseline methods are unable to produce any output.

References

B Amirkhanov, T Nurgazy, G Amirkhanova, M Kunelbayev, and G Tyulepberdinova. Creating 3d models of production equipment and infrastructure using blender. *Int. J. Innovative Res. Sci. Stud.*, 8(1):1572–1588, 2025.

Kiran Bhat, Nishchaie Khanna, Karun Channa, Tinghui Zhou, Yiheng Zhu, Xiaoxia Sun, Charles Shang, Anirudh Sudarshan, Maurice Chu, Daiqing Li, et al. Cube: A roblox view of 3d intelligence. *arXiv preprint arXiv:2503.15475*, 2025.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *ICLR*, 2018.

Aleksey Bokhovkin, Quan Meng, Shubham Tulsiani, and Angela Dai. Scenefactor: Factored latent 3d diffusion for controllable 3d scene generation. In *CVPR*, pages 628–639, 2025.

Ata Çelen, Guo Han, Konrad Schindler, Luc Van Gool, Iro Armeni, Anton Obukhov, and Xi Wang. I-design: Personalized llm interior designer. In *ECCV*, pages 217–234, 2024.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023.

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. In *NeurIPS*, pages 5982–5994, 2022.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. In *NeurIPS*, pages 35799–35813, 2023a.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *CVPR*, pages 13142–13153, 2023b.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Chuan Fang, Yuan Dong, Kunming Luo, Xiaotao Hu, Rakesh Shrestha, and Ping Tan. Ctrl-room: Controllable text-to-3d room meshes generation with layout constraints. In *3DV*, pages 692–701, 2025.

Weixi Feng, Wanrong Zhu, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. In *NeurIPS*, pages 18225–18250, 2023.

Rao Fu, Zehao Wen, Zichen Liu, and Srinath Sridhar. Anyhome: Open-vocabulary generation of structured and textured 3d homes. In *ECCV*, pages 52–70, 2024.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023.

Zeqi Gu, Yin Cui, Zhaoshuo Li, Fangyin Wei, Yunhao Ge, Jinwei Gu, Ming-Yu Liu, Abe Davis, and Yifan Ding. Artiscene: Language-driven artistic 3d scene generation through image intermediary. *CVPR*, pages 2891–2901, 2025.

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

Jiacheng Hong, Kunzhen Wu, Mingrui Yu, Yichao Gu, Shengze Xue, Shuangjiu Xiao, and Deli Dong. Higs: Hierarchical generative scene framework for multi-step associative semantic spatial composition. *arXiv preprint arXiv:2510.27148*, 2025.

Ziniu Hu, Ahmet Iscen, Aashi Jain, Thomas Kipf, Yisong Yue, David A Ross, Cordelia Schmid, and Alireza Fathi. Scenecraft: An llm agent for synthesizing 3d scenes as blender code. In *ICML*, 2024.

Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson WH Lau, Wangmeng Zuo, and Chunchao Guo. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *arXiv preprint arXiv:2506.04225*, 2025a.

Zehuan Huang, Yuan-Chen Guo, Xingqiao An, Yunhan Yang, Yangguang Li, Zi-Xin Zou, Ding Liang, Xihui Liu, Yan-Pei Cao, and Lu Sheng. Midi: Multi-instance diffusion for single image to 3d scene generation. In *CVPR*, pages 23646–23657, 2025b.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023.

Haoran Li, Haolin Shi, Wenli Zhang, Wenjun Wu, Yong Liao, Lin Wang, Lik-Hang Lee, and Pengyuan Zhou. Dreamscene: 3d gaussian-based text-to-3d scene generation via formation pattern sampling. In *ECCV*, pages 214–230, 2024.

Sikuang Li, Chen Yang, Jiemin Fang, Taoran Yi, Jia Lu, Jiazhong Cen, Lingxi Xie, Wei Shen, and Qi Tian. Worldgrow: Generating infinite 3d world. *arXiv preprint arXiv:2510.21682*, 2025.

Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. In *CVPR*, pages 6517–6526, 2024.

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 *CVPR*, pages 300–309, 2023.

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.

Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. In *ICLR*, 2024.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *CVPR*, pages 9970–9980, 2024.

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. In *RSS Workshop DGR*, 2024.

Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, pages 27730–27744, 2022.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. Atiss: Autoregressive transformers for indoor scene synthesis. In *NeurIPS*, pages 12013–12026, 2021.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8026–8037, 2019.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023.

Dustin Podell, Zion English, Kyle Lacey, A. Blattmann, Tim Dockhorn, Jonas Muller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *ArXiv*, abs/2307.01952, 2023.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023.

Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Anselm Levskaya, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. Efficiently scaling transformer inference. In *MLSys*, pages 606–624, 2022.

Hou In Derek Pun, Hou In Ivan Tam, Austin T Wang, Xiaoliang Huo, Angel X Chang, and Manolis Savva. Hsm: Hierarchical scene motifs for multi-scale indoor scene generation. *arXiv preprint arXiv:2503.16848*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.

Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, Zeyu Ma, and Jia Deng. Infinigen indoors: Photorealistic indoor scenes using procedural generation. In *CVPR*, pages 21783–21794, 2024.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, pages 8821–8831, 2021.

Xingjian Ran, Yixuan Li, Linning Xu, Mulin Yu, and Bo Dai. Direct numerical layout generation for 3d indoor scene synthesis via spatial reasoning. *ArXiv*, abs/2506.05341, 2025.

Tianhe Ren and et al. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. *arXiv preprint arXiv:2401.14159*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, pages 36479–36494, 2022.

Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, pages 501–518, 2016.

Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. Lmfusion: Adapting pretrained language models for multimodal generation. *arXiv preprint arXiv:2412.15188*, 2024a.

Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. In *ICLR*, 2024b.

Yawar Siddiqui, Tom Monnier, Filippos Kokkinos, Mahendra Kariya, Yanir Kleiman, Emilien Garreau, Oran Gafni, Natalia Neverova, Andrea Vedaldi, Roman Shapovalov, et al. Meta 3d assetgen: Text-to-mesh generation with high-quality geometry, texture, and pbr materials. In *NeurIPS*, pages 9532–9564, 2024.

Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. In *NeurIPS*, pages 251–266, 2021.

Jiapeng Tang, Yinyu Nie, Lev Markhasin, Angela Dai, Justus Thies, and Matthias Nießner. Diffuscene: Denoising diffusion models for generative indoor scene synthesis. In *CVPR*, pages 20507–20518, 2024.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

UKCMA. Anticipated acquisition by microsoft of activision blizzard, inc. https://www.gov.uk/cmacases/ microsoft-slash-activision-blizzard-merger-inquiry. 2022.

Alexander Vilesov, Pradyumna Chari, and Achuta Kadambi. Cg3d: Compositional generation for text-to-3d via gaussian splatting. *arXiv preprint arXiv:2311.17907*, 2023.

Chunwei Wang, Guansong Lu, Junwei Yang, Runhui Huang, Jianhua Han, Lu Hou, Wei Zhang, and Hang Xu. Illume: Illuminating your llms to see, draw, and self-enhance. In *ICCV*, pages 21612–21622, 2025a.

Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *CVPR*, pages 12619–12629, 2023a.

Xiping Wang, Yuxi Wang, Mengqi Zhou, Junsong Fan, and Zhaoxiang Zhang. Hlg: Comprehensive 3d room construction via hierarchical layout generation. *arXiv preprint arXiv:2508.17832*, 2025b.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *NeurIPS*, pages 8406–8441, 2023b.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *CVPR*, pages 12966–12977, 2025a.

Xiaoyang Wu, Daniel DeTone, Duncan Frost, Tianwei Shen, Chris Xie, Nan Yang, Jakob Engel, Richard Newcombe, Hengshuang Zhao, and Julian Straub. Sonata: Self-supervised learning of reliable point representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22193–22204, 2025b.

Zhennan Wu, Yang Li, Han Yan, Taizhang Shang, Weixuan Sun, Senbo Wang, Ruikai Cui, Weizhe Liu, Hiroyuki Sato, Hongdong Li, et al. Blockfusion: Expandable 3d scene generation using latent tri-plane extrapolation. *ACM TOG*, 43(4):1–17, 2024.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *CVPR*, pages 21469–21480, 2025.

Rongchang Xie, Chen Du, Ping Song, and Chang Liu. Muse-vl: Modeling unified vlm through semantic discrete encoding. In *ICCV*, pages 24135–24146, 2025.

Han Yan, Yang Li, Zhennan Wu, Shenzhou Chen, Weixuan Sun, Taizhang Shang, Weizhe Liu, Tian Chen, Xiaqiang Dai, Chao Ma, et al. Frankenstein: Generating semantic-compositional 3d scenes in one tri-plane. In *SIGGRAPH Asia 2024 Conference Papers*, 2024.

Xiuyu Yang, Yunze Man, Jun-Kun Chen, and Yu-Xiong Wang. Scenecraft: Layout-guided 3d scene generation. In *NeurIPS*, pages 82060–82084, 2024a.

Yandan Yang, Baoxiong Jia, Peiyuan Zhi, and Siyuan Huang. Physcene: Physically interactable 3d scene synthesis for embodied ai. In *CVPR*, pages 16262–16272, 2024b.

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 *CVPR*, pages 16227–16237, 2024c.

Yandan Yang, Baoxiong Jia, Shujie Zhang, and Siyuan Huang. Sceneweaver: All-in-one 3d scene synthesis with an extensible and self-reflective agent. *arXiv preprint arXiv:2509.20414*, 2025.

Shangjin Zhai, Zhichao Ye, Jialin Liu, Weijian Xie, Jiaqi Hu, Zhen Peng, Hua Xue, Danpeng Chen, Xiaomeng Wang, Lei Yang, Nan Wang, Haomin Liu, and Guofeng Zhang. Stargen: A spatiotemporal autoregression framework with video diffusion model for scalable and controllable scene generation. In *CVPR*, pages 26822–26833, 2025.

Qihang Zhang, Chaoyang Wang, Aliaksandr Siarohin, Peiye Zhuang, Yinghao Xu, Ceyuan Yang, Dahua Lin, Bolei Zhou, Sergey Tulyakov, and Hsin-Ying Lee. Towards text-guided 3d scene composition. In *CVPR*, pages 6829–6838, 2024.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *ICLR*, 2025a.

Mengqi Zhou, Xipeng Wang, Yuxi Wang, and Zhaoxiang Zhang. Roomcraft: Controllable and complete 3d indoor scene generation. *arXiv preprint arXiv:2506.22291*, 2025b.

Shijie Zhou, Zhiwen Fan, Dejia Xu, Haoran Chang, Pradyumna Chari, Tejas Bharadwaj, Suya You, Zhangyang Wang, and Achuta Kadambi. Dreamscene360: Unconstrained text-to-3d scene generation with panoramic gaussian splatting. In *ECCV*, pages 324–342, 2024a.

Wenxu Zhou, Kaixuan Nie, Hang Du, Dong Yin, Wei Huang, Siqiang Guo, Xiaobo Zhang, and Pengbo Hu. Il3d: A large-scale indoor layout dataset for llm-driven 3d scene generation. *arXiv preprint arXiv:2510.12095*, 2025c.

Xiaoyu Zhou, Xingjian Ran, Yajiao Xiong, Jinlin He, Zhiwei Lin, Yongtao Wang, Deqing Sun, and Ming-Hsuan Yang. Gala3d: towards text-to-3d complex scene generation via layout-guided generative gaussian splatting. In *ICML*, 2024b.

Junzhe Zhu, Peiye Zhuang, and Sanmi Koyejo. Hifa: High-fidelity text-to-3d generation with advanced diffusion guidance. In *ICLR*, 2024.

Xiaoming Zhu, Xu Huang, Qinghongbing Xie, Zhi Deng, Junsheng Yu, Yirui Guan, Zhongyuan Liu, Lin Zhu, Qijun Zhao, Ligang Liu, et al. Imaginarium: Vision-guided high-quality 3d scene layout generation. *arXiv preprint arXiv:2510.15564*, 2025.