ControlRoom3D: Room Generation using Semantic Proxy Rooms

Jonas Schult^{1,2*} Sam Tsai^{1} Lukas Höllein^{1,3*} Bichen Wu^{1} Jialiang Wang^{1} Chih-Yao Ma^{1} Kunpeng Li^{1}
Xiaofang Wang^{1} Felix Wimbauer^{1,3*} Zijian He^{1} Peizhao Zhang^{1} Bastian Leibe^{2} Peter Vajda^{1} Ji Hou^{1}
^{1}Meta GenAI ^{2}RWTH Aachen University ^{3}Technical University of Munich

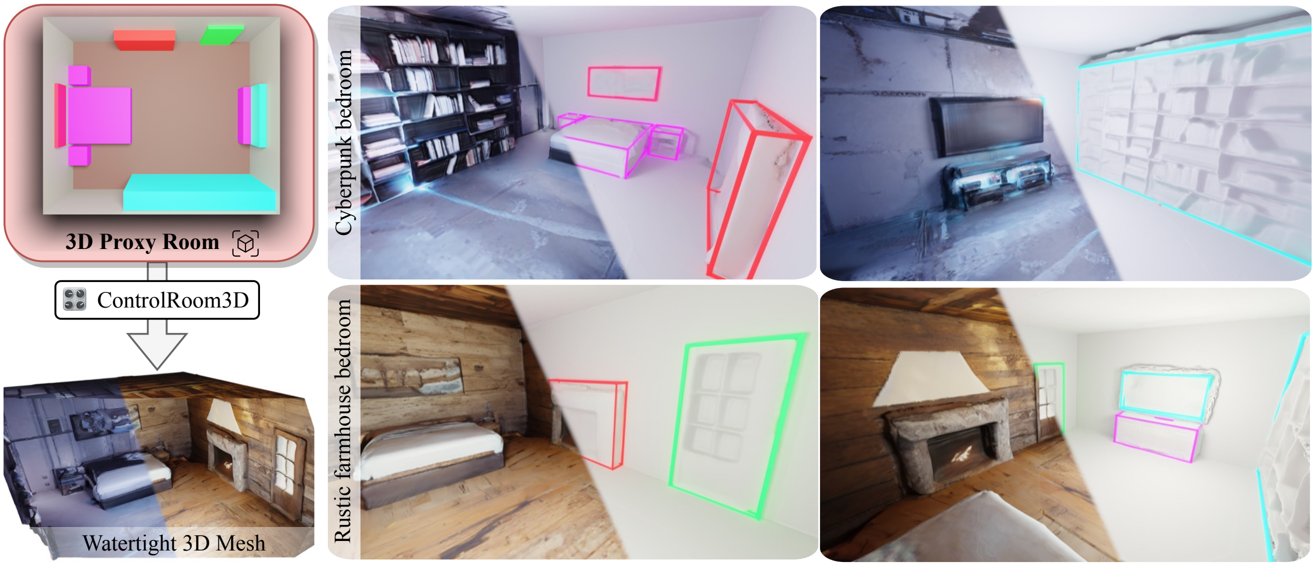

Figure 1. **Textured 3D Mesh Generation from 3D Semantic Layouts and Text Prompts.** Given a textual description of the overall room style and a rough 3D room layout based on 3D semantic bounding boxes, our method called ControlRoom3D creates diverse and globally plausible 3D room meshes which align well with the room layout.

Abstract

Manually creating 3D environments for AR/VR applications is a complex process requiring expert knowledge in 3D modeling software. Pioneering works facilitate this process by generating room meshes conditioned on textual style descriptions. Yet, many of these automatically generated 3D meshes do not adhere to typical room layouts, compromising their plausibility, e.g.*, by placing several beds in one bedroom. To address these challenges, we present* ControlRoom3D*, a novel method to generate high-quality room meshes. Central to our approach is a user-defined 3D semantic proxy room that outlines a rough room layout based on semantic bounding boxes and a textual description of the overall room style. Our key insight is that when rendered to 2D, this 3D representation provides valuable geometric and semantic information to control powerful 2D models to generate 3D consistent textures and geometry that aligns well with the proxy room. Backed up by an extensive study including quantitative metrics and qualitative user evaluations, our method generates diverse and globally plausible 3D room meshes, thus empowering users to design 3D rooms effortlessly without specialized knowledge.*

1. Introduction

Generating high-quality mesh representations of 3D scenes is essential for creating immersive experiences in AR/VR applications. However, manually synthesizing these 3D environments is a complex task that requires expertise in 3D modeling software [1, 17], posing a substantial barrier for end-users looking to create personalized scenes. Fueled by the recent progress in 2D generative models [13, 19, 39, 45], an emerging line of work aims to simplify the creation process of 3D objects [9, 26, 48, 52, 53] or even room-scale scenes [10, 15, 20, 57]. Among them, the recent work of Text2Room [20] creates compelling textured 3D room-scale meshes using a text prompt describing the scene and an off-the-shelf text-to-image model [39]. This is achieved by iteratively inpainting missing 2D regions of the rendered 3D mesh from a novel viewpoint. Subsequently, depth is estimated for the new content, followed by backprojection and fusion with the current 3D mesh. Despite Text2Room's ability to generate locally convincing meshes with detailed textures, it suffers from two crucial shortcomings due to its iterative mesh inpainting strategy solely based on local context: First, Text2Room does not follow conventional room layouts, yielding unrealistic spatial arrangements of objects and walls. Secondly, it tends to duplicate the most promi-

---

[*] Work performed during an internship at Meta GenAI.

nent objects of a scene type, e.g., placing several beds into a single bedroom, leading to semantically implausible scenes.

To address these shortcomings, we propose Control-Room3D, which generates diverse and globally plausible 3D room meshes. A central element of our method is the 3D semantic proxy room (Fig. 1) which allows the user to specify semantic bounding boxes and textual descriptions of the desired room style that then serve as a rough outline for the generated room layout. Our key insight is that this 3D representation, when rendered to 2D, provides valuable geometric and semantic information. This information can be used to control powerful 2D generative models, enabling the creation of 3D consistent textures and geometry that align with the proxy room. However, projecting 3D controls to 2D comes with a loss of information which needs to be properly addressed. As naively generating 2D views individually is detrimental for scene style consistency, we first present a novel panorama generation method, which generates a comprehensive 360° view of the room. Leveraging homographic constraints between equirectangular patches, we equip a latent diffusion model with correspondence-aware attention modules [49] which establish a coherent and seamless style, while faithfully adhering to the projected semantics of the user-provided 3D proxy room. Next, the quality of depth estimation becomes pivotal for converting this panorama into a 3D representation that aligns with the proxy room. Building upon our insight that the 3D proxy room offers not only semantic but also valuable geometric information, we introduce a novel geometry alignment approach which leverages the spatial dimensions of 3D bounding boxes in an iterative optimization strategy to support a metric depth estimation network [6] to generate 3D structures which align with the proxy room. Thirdly, while panoramic views are valuable, they are insufficient for a complete 3D scene. To address issues like occlusion and low-resolution textures, our method first eliminates low density regions, then uses new viewpoints for inpainting, guided by the proxy room and our geometry alignment, thus integrating new geometry into the mesh seamlessly.

To summarize, our contributions are: (1) We propose ControlRoom3D which generates diverse and semantically plausible 3D room meshes aligning well with the user-defined room layout and textual description of the room style. (2) Leveraging our key insight that this room layout provides both valuable semantic and geometric priors, we present technical components including the semantic proxy room, guided panorama generation, mesh completion, and geometry alignment to enable coherent and seamless 3D textures and geometry. (3) Using both quantitative and qualitative metrics, we validate the effectiveness of each component introduced and show that our method generates more plausible 3D rooms compared to competing methods.

2. Related Work

Text-to-3D. Fueled by the availability of large-scale vision-text datasets [21, 31, 42–44, 51, 56] and powerful vision-language models [34], recent advancements in 2D diffusion models [5, 11, 30, 35, 39, 40] have significantly contributed to the ongoing progress in generating 3D objects and scenes. In particular, one prominent research direction focuses on single object synthesis [26, 29, 33, 48, 53, 57], wherein a differentiable 3D representation [22, 28] is commonly optimized through a loss signal derived from the denoising of rendered views [33, 52, 53]. Despite achieving impressive results in the synthesis of object-centric scenes, the challenge of generalizing these techniques to large and complex scenes persists [53]. Typically, existing approaches rely on additional 3D data sources [3, 4, 12], which are challenging to obtain at scale. Two notable approaches related to our work include SceneScape [15] and Text2Room [20]. These methods leverage off-the-shelf text-conditioned inpainting models [39] in combination with state-of-the-art monodepth estimation models [2, 36] to create a 3D mesh representation. Alternatively, LDM3D [47] directly predicts an RGBD frame based on a given text prompt. However, these approaches have limited control over the generation process by only being able to change the text prompt at each inpainting step, typically leading to globally implausible room layouts. In contrast, ControlRoom3D allows precise control over the scene composition by using globally consistent 2D maps derived from 3D semantic proxy rooms to guide the generation process.

Semantically Guided 3D Scene Generation. An emerging line of works investigates scene generation with semantic compositing [8, 14, 16, 25, 39, 54, 55]. However, in 3D they are typically constrained to low resolution textures and partial room settings to enable inward-facing views [10, 32, 57] or require synthetic 3D datasets and thus are restricted to the room style present in the dataset [3]. At the same time, current diffusion models show great capabilities to be extended by control signals such as pixel-perfect maps, key-points or bounding boxes [8, 25, 54, 55, 58]. We follow this line of work and show that 2D control signals derived from a 3D representation can be used to produce 3D consistent high-resolution images of large diversity without the need of inward-facing views.

Mesh Texturization. A related line of work deals with re-texturizing a given mesh representation of 3D scenes [46, 49] or objects [27] using a stylization text prompt. Unlike ControlRoom3D, these methods focus on re-texturizing the mesh in a novel style without (or only minor adaptations) to the original geometry while our approach not only generates stylized textures but also creates novel geometry from scratch which aligns with the style of the texture.

Figure 2. Method Overview. The 3D semantic proxy room 🟥 contains a user-defined 3D room layout based on semantic bounding boxes and a textual description of the overall room style. From this representation, we derive control signals that ensure 3D consistent textures and align corresponding depth estimates with the room layout. The panorama generation module 🟦 generates a comprehensive 360° view of the room, the depth of which is then estimated and aligned with the proxy room 🟩. After Poisson reconstruction and artifact cleaning 🟪, we update the room mesh by iteratively completing partial renderings of the scene 🟨 and estimating the depth of newly added textures 🟩.

3. Method

Overview. (Fig. 2) Given a textual description of the overall room style and a rough 3D room layout based on 3D semantic bounding boxes, ControlRoom3D creates diverse and globally plausible 3D room meshes which align well with the room layouts. The key technical components of our method comprise the *proxy room* 🟥, from which we derive 2D semantic and guidance depth maps for arbitrary viewpoints in the scene; the *geometry alignment module* 🟩, to align newly generated 2D views to the proxy room; the *panorama generation module* 🟦, to initially generate a globally style-consistent 360° view of the scene; and a final *completion module* 🟨, which samples novel camera viewpoints to inpaint occluded and unobserved regions in the mesh with high-resolution textures. Details of each of these steps are discussed in the following sections, and we highlight each component with its corresponding color of Fig. 2.

3D Semantic Proxy Room. (Fig. 2, 🟥) At the core of our method lies the user-defined 3D semantic proxy room, which acts as a rough outline for the room layout and is defined by semantic bounding boxes and a textual description of the desired room style. Specifically, we define a proxy room M as a set of semantic bounding boxes $B = (p, s, c, i)$, where $p \in \mathbb{R}^3$ is the box center, $s \in \mathbb{R}^3$ the box size, c the semantic class id, and i the unique instance id associated with each box. Our key insight is that this 3D representation, when rendered to 2D, provides valuable geometric and semantic information that can be used to control powerful 2D generative models, enabling the creation of 3D consistent textures and geometry that align with the proxy room. We render the proxy scene from a given viewpoint V, producing multiple types of frames:

$$S, I, D_n, D_f = r(M, V)$$

Here, $r(\cdot)$ is standard rasterization without shading, the outputs $S \in [1, C]^{H \times W}$, $I \in \mathbb{N}^{H \times W}$, $D_n, D_f \in \mathbb{R}_+^{H \times W}$ are the rendered 2D semantic segmentation map of C classes, the instance segmentation map, and the near and far guidance

Figure 3. Proxy Room Renderings. We use classical rasterization to render semantic segmentation and instance maps as well as near and far depth maps from the proxy room.