**ControlRoom3D: Room Generation using Semantic Proxy Rooms**

Jonas Schult*, Sam Tsai, Lukas Hölein*, Bichen Wu, Jialiang Wang, Chih-Yao Ma, Kunpeng Li, Xiaofang Wang, Felix Wimbauer*, Zijian He, Peizhao Zhang, Bastian Leibe, Peter Vajda, Ji Hou

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.

---

* Work performed during an internship at Meta GenAI.

---

**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-
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 ControlRoom3D, 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, keypoints 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.
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 \( \mathcal{M}_v \) as a set of semantic bounding boxes \( \mathcal{B} = (p, s, c, i) \), 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:

\[
S, I, D_n, D_f = r(\mathcal{M}_v, V)
\]

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
depth map, respectively (Fig. 3). In the following, we detail how these rendered maps are utilized in our components to achieve room mesh generation that aligns both semantically and geometrically with the proxy room.

Guided Panorama Generation. (Fig. 2, and Fig. 4) We generate 360° panoramas guided by 3D semantic bounding boxes rendered from the proxy room. To this end, we equip an LDM with correspondence-aware attention modules (CAA) [49] and semantic and depth adapters finetuned on semantic bounding boxes.

Geometry Alignment. (Fig. 2, and Fig. 5) After generating panoramas, the next step is to transform these 2D textures into a 3D representation by predicting per-pixel depth. Naively estimating depth using a standard depth estimator can yield subpar results due to scale ambiguity. However, the proxy room offers valuable geometric information, including the near and far guidance depth map (Fig. 3c,d). A fundamental aspect of ControlRoom3D is to align the predicted depth with this geometric prior. Specifically, we optimize the depth predicted by a pre-trained metric depth estimator (MDE) [6] to fit within its respective bounding boxes while preserving the object’s shape (Fig. 5). To achieve this, we introduce a geometry alignment loss:

$$\mathcal{L}_d = \begin{cases} 0 & \text{if } D_n \leq \hat{d} \leq D_f, \\ \min \left( \|\hat{d} - D_n\|_1, \|\hat{d} - D_f\|_1 \right) & \text{otherwise.} \end{cases}$$

Adapter Fine-tuning on HyperSim Bounding Boxes. Off-the-shelf semantic adapters gain strong semantic understanding through training on large datasets, but they are typically trained on pixel-perfect semantic masks. This stands in stark contrast to our control input, which involves 3D bounding boxes rendered in 2D views that only provide a rough outline of the object (Fig. 3a). To bridge this gap, we construct a dataset derived from HyperSim [38], containing photo-realistic renderings of indoor environments with 3D semantic bounding box annotations. We render 3D bounding boxes for each scene and fine-tune the semantic and depth adapters on this dataset. Additional details can be found in our supplementary material.
Here, \( \hat{d} \) is the predicted depth, and \( D_n, D_f \) denote the near and far depth values. Note that for walls, ceiling, and floor \( D_n = D_f \). This loss penalizes depth predictions that fall outside the designated bounding box with an \( L_1 \) loss. Additionally, we employ SAM [23] to extract pixel-perfect instance masks for each generated instance by prompting it with the bounding box extracted from the instance maps \( I \) (Fig. 3b). For pixels within the rendered bounding box but outside the SAM mask, we set \( D_n \) to \( D_f \). During test time, we optimize a pre-trained MDE using this loss, thereby aligning the depth with the proxy room while retaining the depth estimator’s strong prior learned from large datasets. While the depth alignment loss \( \mathcal{L}_d \) aligns the frame with the proxy room, it may alter the object’s surface significantly while fitting it into its bounding box. To mitigate this, we present a normal preservation loss: \( \mathcal{L}_n = \| \hat{n} - n_0 \|_1 \) where \( \hat{n} \) are the estimated normals from the current depth prediction \( \hat{d} \) and \( n_0 \) are the normals from the original MDE.

Once we have predicted depth values that align with the proxy room, our next step is to convert the 2D textures into a 3D representation. However, a straightforward unprojection using the depth values can result in visible seams in overlapping regions, because the depth values for each frame are individually optimized. To address this challenge, we first convert the z-depth values into distance values, as they remain invariant to rotations. We then combine all frames into a panoramic view using spherical projections scaled by their corresponding distance values followed by pyramid blending, ultimately yielding the initial 3D representation.

**Mesh Cleaning.** (Fig. 2, \( \square \) and Fig. 6) While RGBD panoramas offer an initial 3D representation, they are susceptible to issues like noisy depth estimates and partially low-resolution textures, particularly in areas that are occluded or where mesh faces are stretched. To address these challenges, our approach begins by identifying and filtering out regions of low quality in the mesh. We then synthesize new content for these regions from a more favorable viewing angle. To this end, we first perform Poisson surface reconstruction, converting the 3D point cloud obtained from the RGBD panorama into a watertight mesh representation \( \mathcal{M}_b \) (Fig. 6a). We proceed to remove mesh vertices and their adjacent faces if their density falls below a specified threshold \( \delta_c \). Finally, we eliminate floating artifacts that are disconnected from the main mesh by deleting components with a vertex count below a specified threshold \( \delta_v \) (Fig. 6b).

**Mesh Completion.** (Fig. 2, \( \square \) and Fig. 7) After Poisson reconstruction and artifact removal, we further refine the room mesh by iteratively inpainting the scene from various favorable viewpoints \( V \), effectively filling missing areas in the mesh with novel content. To this end, we render the current 3D mesh \( \mathcal{M}_b \) with classical rasterization \( r(\cdot) \) to obtain the RGB frame \( F \), depth map \( D_b \) as well as the image mask \( M \) that identifies pixels without observed content: \( F, D_b, M = r(\mathcal{M}_b, V) \). In order to unleash the full potential of powerful 2D generation models, which perform better with larger connected regions, we first inpaint small gaps using traditional Telea inpainting [15, 20, 50]. We then expand the remaining holes through a sequence of morphological operations, specifically erosion followed by dilation. This modified mask, combined with the rendered 2D semantic \( S \) and guidance depth map \( D_n \) from the proxy room, guides the inpainting of the missing areas using a semantic and depth adapter, yielding the final generated RGB frame.

We integrate this new content into the mesh using our geometry alignment module (Fig. 2, \( \square \)). For seamless integration with pre-existing structures, we superimpose the rendered depth \( D_b \) on the near and far guidance depth map \( D_n, D_f \) to optimize the depth of novel content to correspond to already generated 3D structures, i.e., \( D^i_0 = D^i_1 = D^i_f \) for \( M^i_j \). We repeat this process for a number of selected viewpoints, effectively inpainting any missing region while seamlessly integrating novel content.
4. Experiments

Evaluation Setup and Metrics. We use distinct room layouts encompassing typical indoor rooms such as bedroom or kitchen but also unique settings such as wellness spa with an indoor pool as targeted room for evaluation. For text prompting, we define 12 room styles, spanning a diverse range of geographical (e.g., Japanese, Scandinavian), seasonal (e.g., Halloween, Christmas), historical (e.g., Ancient Roman, Stone-age cave), and artistic (e.g., cyberpunk, industrial) themes. We ask 12 participants to rate these scenes on a scale of 1–5 with respect to the following dimensions: 3D structure completeness (3DS), proxy alignment to the 3D guidance (PA), generated room layout plausibility (LP), and overall perceptual quality based on user preference (PQ). We render 25sec long videos with 650 frames each for each generated room showing all objects from multiple (challenging) angles. We let users rate each scene individually (no head-to-head comparison). In total, we collected 1120 data points using 175 scenes from the user study. We provide further details about the user study in the supplementary material. Moreover, we calculate the CLIP score (CS) [18] and Inception score (IS) [41] on 100 RGB renderings from random viewpoints for each scene.

Implementation Details. We use a pre-trained LDM [39] as well as semantic and depth adapter modules. We finetune both the adapter weights on our 3D bounding box dataset derived from the HyperSim [38] dataset using BLIPv2 captions [24]. The correspondence-aware attention module (CAA) is trained on the Matterport3D [7] dataset. Our metric depth estimator is a pre-trained ZoeDepth [6] model. We use PyTorch3D [37] for mesh rasterization and fusion for the 3D proxy room as well as the final generated mesh.

4.1. Comparison with State-of-the-Art Methods

Baselines. There are no directly comparable methods which generate 3D textured meshes of entire rooms given a 3D room layout and a textual style description. We thus adapt top-performing related methods for this task. Text2Room [20] uses a text-conditioned inpainting model followed by mono-depth inpainting [2]. Following their original instructions, we enforce a room layout by changing the text prompt during the camera trajectory to reflect which objects are visible in the current frame. MVDiffusion [49] creates panorama images by enforcing correspondences between overlapping image crops using homography constraints. Here, for each image crop we adjust the text condition to reflect which objects should be visible and use the same off-the-shelf mono-depth estimator [6] and poisson mesh reconstruction as our method.

Results. In our study, we provide a detailed quantitative comparison with top-performing methods in Tab. 1 and illustrate qualitative examples in Fig. 8. A critical finding is that text-only scene conditioning is inadequate for generating realistic room layouts. Instead, our method integrates 2D semantic and geometric controls based on a 3D proxy room leading to more plausible layouts with an improvement of +1.5 LP. For instance, Text2Room frequently but incorrectly places multiple instances of the most prominent object of a scene type in the room, such as 4 beds in a single bedroom (Fig. 8, bottom), resulting in unrealistic layouts. In contrast, our approach accurately reflects the intended layout, marked by colored bounding boxes on the renderings. This accuracy enables more efficient completion of missing areas reducing artifacts (+1.7 3DS) and enhancing overall perceptual quality (+1.2 PQ) while faithfully following the text conditioning (+2.8 CS). The inception score (IS) tends to favor methods that generate diverse scenes, even if they do not necessarily represent plausible room layouts (LP).

Nonetheless, our method displays versatility in creating a wide range of scene types and layouts, as shown in Fig. 8 and supplementary videos.

4.2. Analysis Experiments

The key technical components of our methods comprise the proxy room, the geometry alignment module, the panorama generation, and the final mesh completion. In Tab. 2, we evaluate the influence of each component. In the supplementary material, we provide further results about the adapter fine-tuning on HyperSim as well as visualizations of baseline methods of the ablation study.
Figure 8. **Qualitative Comparison with Baseline Methods.** We show colored geometry renderings of our method and two baseline methods. ControlRoom3D generates convincing geometries and textures given a user-defined room layout and a textual style description. We provide further example images and videos in the supplementary material.
Figure 9. **Fine-tuning with rendered bounding boxes from HyperSim.** Prior to fine-tuning with HyperSim data, ControlRoom3D tends to create box-shaped objects and struggles to accurately generate objects within their respective bounding boxes (a), whereas fine-tuning enables ControlRoom3D to fill 3D bounding boxes with appropriate content (b).

**Adapter Fine-tuning on HyperSim Bounding Boxes.** Fig. 9 presents a side-by-side qualitative evaluation of ControlRoom3D, illustrating its performance with and without fine-tuning on our 3D bounding box dataset derived from HyperSim. A critical observation is that when ControlRoom3D uses adapters only trained on datasets of pixel-precise masks, it faces challenges in generating objects accurately within their designated bounding boxes. Instead, it tends to create box-shaped objects that entirely fill the bounding boxes (Fig. 9a). However, when the adapters are fine-tuned using our dataset derived from HyperSim, ControlRoom3D demonstrates an improved ability to generate objects that appropriately fit within their corresponding bounding boxes (Fig. 9b).

**Consistent Style Panorama Generation.** The participants in our user study expressed a strong preference for the panorama generation module (Fig. 10a), noting its superiority across all three metrics compared to the incremental inpainting approach. Incremental inpainting involves warping generated images to subsequent ones and autoregressively filling in unseen areas. Since this inpainting is solely based on local context, it leads to gradual changes in style and challenges in achieving accurate loop closure [49] (Fig. 10a, white box). In contrast, our panorama generation module creates a complete 360° panorama, conditioned on both semantic and geometric 3D information in a unified manner. This approach ensures a consistent style throughout the entire scene, effectively addressing the issues associated with incremental inpainting and resulting in a more cohesive and visually appealing overall scene (Fig. 10b, bottom).

**Geometry Alignment with 3D Boxes.** Estimating the exact scale of a 3D scene from a single RGB image is an ill-posed task, often leading to severe inaccuracies in even the most advanced metric depth estimators [6]. As shown in Fig. 11, without iterative alignment, the estimated depth would lead to implausible geometry (t=1). By iteratively optimizing the 3D representation with additional geometric information of 3D bounding boxes of the proxy room layout, the final geometry is more plausible (t=600). The geometry alignment module is the core ingredient to achieve strong alignment with the proxy room, as evident by the substantial improvement of +2.32 PA in Tab. 2.

In Fig. 12, we present an additional qualitative ablation study focusing on the use of SAM masks and the normal preservation loss, both integral components of the depth alignment module. We leverage SAM [23] to obtain pixel-precise instance masks for each generated object. For pixels located within the rendered bounding box but outside the SAM mask, we assign the depth value $D_n$ to $D_f$. As shown in Fig. 12 (top), including SAM masks leads to sharper 3D...
object boundaries, resulting in a more seamless integration into the 3D room mesh. Although the depth alignment loss $L_d$ effectively aligns the frame with the 3D proxy room, it may occasionally distort the surface of objects to fit them within their bounding boxes. To counter this, we introduce the normal preservation loss $L_n$, retaining the original shape of the objects (Fig. 12, bottom).

High Quality Generation with Mesh Completion. Once the geometry alignment module has aligned the generated geometry with the layout, we can render control signals from the proxy room, including semantic, instance, and guidance depth maps, for selected new viewpoints that correspond with the created 3D mesh. This process enables us to fill in the missing parts of the mesh which remained unobserved after the panorama generation with localized inpainting, and incorporate new content into the 3D structure, guided by these aligned control signals. This yields notable enhancement in structural completeness (+0.54 3DS), culminating in the highest perceptual quality (4.23 PQ).

5. Conclusion
In this paper, we proposed ControlRoom3D, a novel method to generate high-quality 3D room meshes given a textual description of the room style and a user-defined room layout outlined by 3D semantic bounding boxes. When rendered to 2D, the geometric and semantic information is used to guide generative models to create texture and geometry aligned with the proxy room’s layout. Naive guidance does not produce plausible meshes, thus, we introduced guided panorama generation and geometry alignment modules to ensure consistent style and plausible geometry. Finally, we introduced mesh post-processing and completion methods, improving the final generation quality substantially. We believe our method marks a significant step forward in enabling the creation of large-scale 3D assets through user-defined control signals, greatly simplifying the process for end-users without any expert knowledge in 3D modeling software to design personalized scenes.

Acknowledgments. We would like to thank Xiaoliang Dai, Jia-Bin Huang, Jeff Liang and Matthew Yu for helpful feedback and discussions.
References

[20] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2Room: Extracting Textured 3D Meshes from 2D Text-to-Image Models. In ICCV, 2023. 1, 2, 5, 6, 12
[21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In ICML, 2021. 2
[24] Junning Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In ICML, 2023. 6
[25] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. GLIGEN: Open-Set Grounded Text-to-Image Generation. CVPR, 2023. 2
ControlRoom3D: Room Generation using Semantic Proxy Rooms

Supplementary Material

In this supplementary material, we provide mode details of our technical components and experimental setup. Specifically, we provide further details about our user study and conclude with additional qualitative results. For a comprehensive understanding, we suggest that reviewers watch our supplemental video, which includes detailed explanations and showcases videos of 3D room meshes in a variety of room types and styles.

6. Adapter Fine-tuning on HyperSim

Dataset Preparation and Implementation Details. The HyperSim dataset [38] offers 2D rendered images that include camera positions and aligned 3D semantic bounding boxes. Using these camera poses, we project the semantic bounding boxes into 2D, creating depth maps and semantic maps of the bounding boxes (see Fig. 13, top). However, we observed that some images from the HyperSim dataset do not meet our quality requirements (Fig. 13, bottom). Therefore, we exclude images where the camera roll exceeds ±8.6°, as our camera setup is typically parallel to the ground. Additionally, we discard images where a single semantic class covers more than half of the image area, as this often suggests the camera is too close to objects in the 3D scene or inside a bounding box. Images with a maximum depth value over 15 meters are also removed, considering our focus on bounded indoor environments. For the selected images, we calculate BLIPv2 captions. We fine-tune the semantic and depth adapter independently on their respective maps. We train for 5000 iterations on 8 A100 GPUs, with a batch size of 8 and a learning rate of 1e-7.

Qualitative Comparison. Fig. 9 presents a side-by-side qualitative evaluation of ControlRoom3D, illustrating its performance with and without fine-tuning on our 3D bounding box dataset derived from HyperSim. A critical observation is that when ControlRoom3D uses adapters only trained on datasets of pixel-precise masks, it faces challenges in generating objects accurately within their designated bounding boxes. Instead, it tends to create box-shaped objects that entirely fill the bounding boxes (Fig. 9a). However, when the adapters are fine-tuned using our dataset derived from HyperSim, ControlRoom3D demonstrates an improved ability to generate objects that appropriately fit within their bounding boxes (Fig. 9b).

7. Details on User Study

We carried out two user studies in which we ask 12 participants to rate 3D scenes on a scale of 1–5 with respect to three qualitative dimensions. Firstly, we compare our work with related work, i.e., Text2Room [20] and an adaption of MVDiffusion [49] for 3D room generation. We ask all participants to rate each scene individually with respect to 3D structure (3DS), i.e., “is the 3D mesh complete and not distorted?”, layout plausibility (LP), i.e., “does the room layout resemble a typical room layout of the specified type?” and overall perceptual quality (PQ). We show the user study interface in Fig. 14. Secondly, we conduct a user study for the ablation study in which we evaluate the influence of each technical component to our full model. Here, we replace the question about layout plausibility with proxy alignment (PA), i.e., “are objects generated within their corresponding bounding box?”, and additionally superimpose 3D bounding boxes on the video to visualize the proxy room guidance. An example of this is included in the supplementary material.

Figure 13. HyperSim Filtering. We show samples rendered from the HyperSim dataset which meet our quality requirements (top). Samples with camera roll, extreme far shots and close-ups are filtered out to better match ControlRoom3D’s use case (bottom).
We render 25sec long videos for each generated room showing all objects from multiple (challenging) angles. We ask participants to rate these scenes on a scale of 1–5 with respect to the following dimensions: 3D structure completeness (3DS), proxy alignment to the 3D guidance (PA), generated room layout plausibility (LP), and overall perceptual quality (PQ).

Figure 15. Qualitative Ablation on the key components of our method. We notice that 2D guidance (c) plays a crucial role in creating a plausible room layout, cf. (a) – (b). Nevertheless, the scene still lacks geometric alignment with the proxy room (cf. transparent bounding box overlays). Our depth alignment module accurately aligns the generated scene with the proxy room geometry (d). To address bad reconstruction artifacts due to unobserved areas (bottom left), we employ our semantic completion stage to inpaint these missing regions. This results in complete 3D rooms without blurred-out sections (bottom right).

8. Further Qualitative Results

Visualization of the Ablation Study. We provide an additional qualitative study regarding the key technical components of ControlRoom3D in Fig. 15 and show videos of an exemplary 3D room of each ablation study experiment in the supplementary material.

Additional qualitative results of our main method. In Fig. 16, we show additional qualitative examples of ControlRoom3D. Moreover, we recommend that reviewers watch our supplemental video explaining main technical component of ControlRoom3D as well as qualitative videos of 3D room meshes of various room types and styles.
Figure 16. **Further Qualitative Examples.** We show colored geometry renderings of our method. ControlRoom3D generates convincing geometries and textures given a user-defined room layout and textual style description. More qualitative videos can be found at the end of the explanatory video in the supplementary material.