ControlRoom3D: Room Generation using Semantic Proxy Rooms

Supplementary Material

In this supplementary material, we provide mode details of our technical components and experimental setup. Moreover, we show a qualitative ablation for the depth alignment module focusing on the integration of SAM masks as well as the normal preservation loss. We then provide further details about our user study and conclude with additional qualitative results. For a comprehensive understanding, we suggest to watch the supplemental video provided on our project page[1]*, which includes detailed explanations and showcases videos of 3D room meshes in a variety of room types and styles.*

1. Adapter Fine-tuning on HyperSim

Dataset Preparation and Implementation Details. The HyperSim dataset [2] 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. 1, *top*). However, we observed that some images from the HyperSim dataset do not meet our quality requirements (Fig. 1, *bottom*). Therefore, we exclude images where the camera roll exceeds $\pm 8.6°$, 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. 3 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. 3a). 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. 3b).

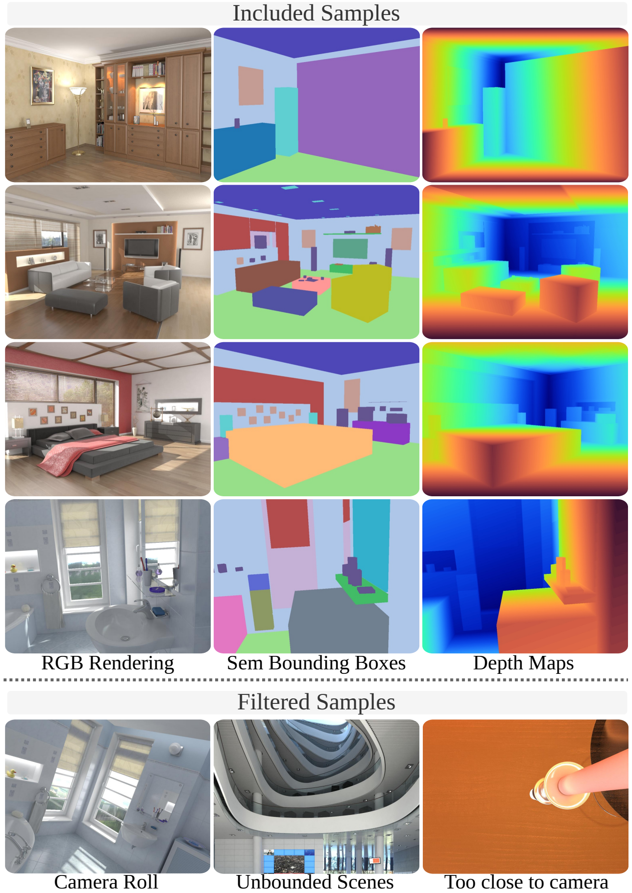


Figure 1. **HyperSim Filtering.** We show samples rendered from the HyperSim dataset which meet our quality requirements (*top*). Samples with camera roll, extreme far shots and close-ups are filtered out to better match ControlRoom3D's use case (*bottom*).

2. Qualitative Ablation on Depth Alignment

Using a standard metric depth estimator can lead to inaccurate results due to scale ambiguity. The proxy room, however, provides essential geometric data, including the near and far guidance depth maps, which are utilized in our depth alignment module, as detailed in the paper. In Fig. 2, 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 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. 2 (*top*), including SAM masks leads to sharper 3D object boundaries, resulting in a more seamless integration into the 3D room

[1] Project Page: jonasschult.github.io/ControlRoom3D

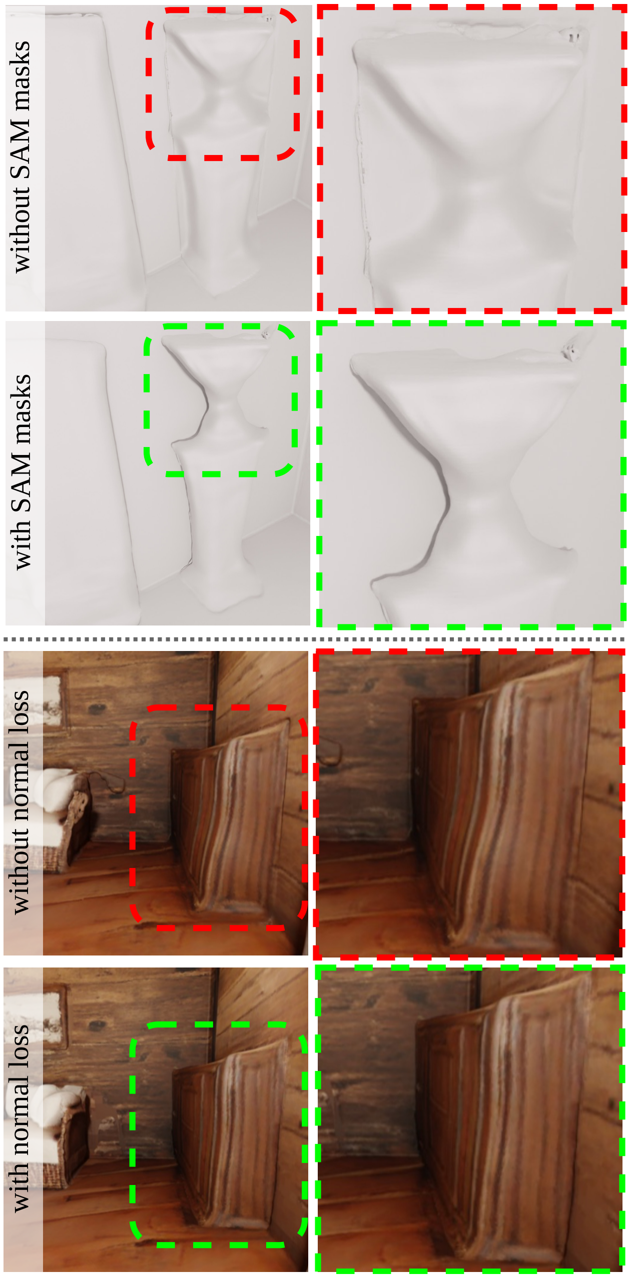


Figure 2. **Ablation on Depth Alignment.** Leveraging pixel-perfect SAM masks yields crisper 3D object boundaries (*top*). While the depth alignment loss aligns the frame with the 3D proxy room, it can distort the object's surface as it fits the depth within its bounding box. Addressing this issue, we introduce our normal preservation loss to maintain the object's original shape (*bottom*).

mesh. Although the depth alignment loss $\mathcal{L}_d$ effectively aligns the frame with the 3D proxy room, as illustrated in Fig. 10 in our paper, it may occasionally distort the surface of objects to fit them within their bounding boxes. To counter this, we introduce the normal preservation loss $\mathcal{L}_n$, retaining the original shape of the objects (Fig. 2, *bottom*).

3. 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 [1] and an adaption of MVDiffusion [3] 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. 4. 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.

4. Further Qualitative Results

Visualization of the Ablation Study. We provide an additional qualitative study regarding the key technical components of ControlRoom3D in Fig. 5.

Additional qualitative results of our main method. In Fig. 6, we show additional qualitative examples of ControlRoom3D. Moreover, we recommend to watch the supplemental video provided on our project page explaining main technical component of ControlRoom3D as well as qualitative videos of 3D room meshes of various room types and styles.

References

[1] 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. 2

[2] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A Photorealistic Synthetic Dataset for Holistic Indoor Scene Understanding. In *ICCV*, 2021. 1

[3] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. MVDiffusion: Enabling Holistic Multi-view Image Generation with Correspondence-Aware Diffusion. In *NeurIPS*, 2023. 2

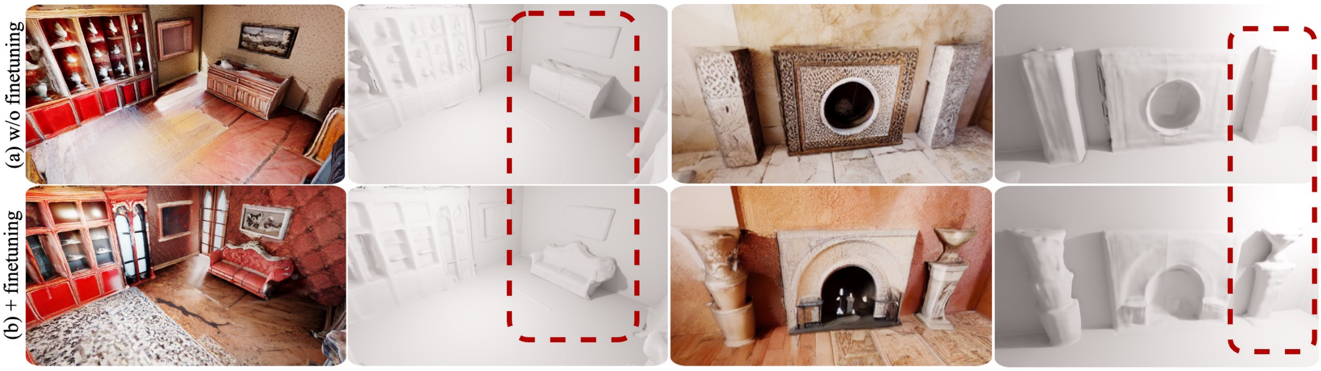
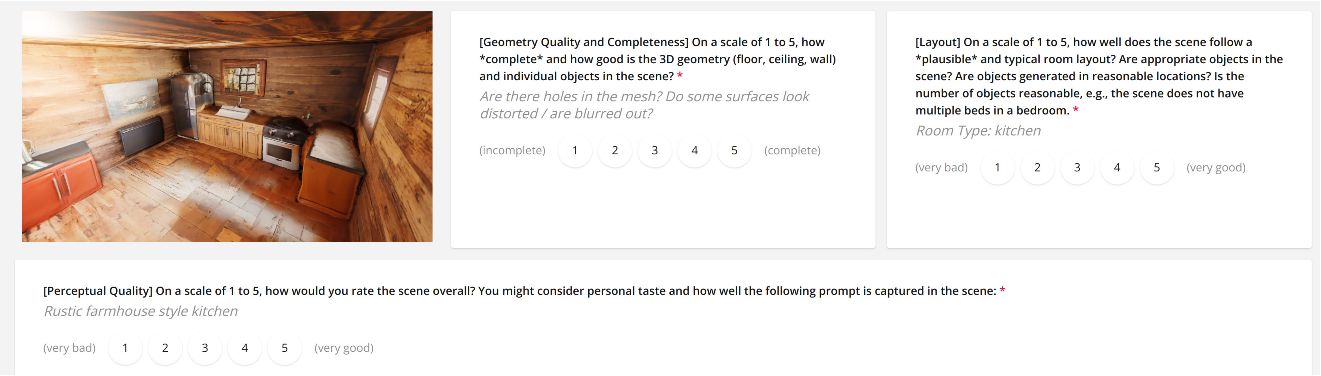


Figure 3. **Fine-tuning with rendered bounding boxes from HyperSim.** Prior to fine-tuning with HyperSim data, ControlRoom3D tends to create box-shaped objects and struggles to accurately generate objects within their respective bounding boxes (*a*), whereas fine-tuning enables ControlRoom3D to fill 3D bounding boxes with appropriate content (*b*).



[Geometry Quality and Completeness] On a scale of 1 to 5, how *complete* and how good is the 3D geometry (floor, ceiling, wall) and individual objects in the scene? *

Are there holes in the mesh? Do some surfaces look distorted / are blurred out?

(incomplete) ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 (complete)

[Layout] On a scale of 1 to 5, how well does the scene follow a *plausible* and typical room layout? Are appropriate objects in the scene? Are objects generated in reasonable locations? Is the number of objects reasonable, e.g., the scene does not have multiple beds in a bedroom. *

Room Type: kitchen

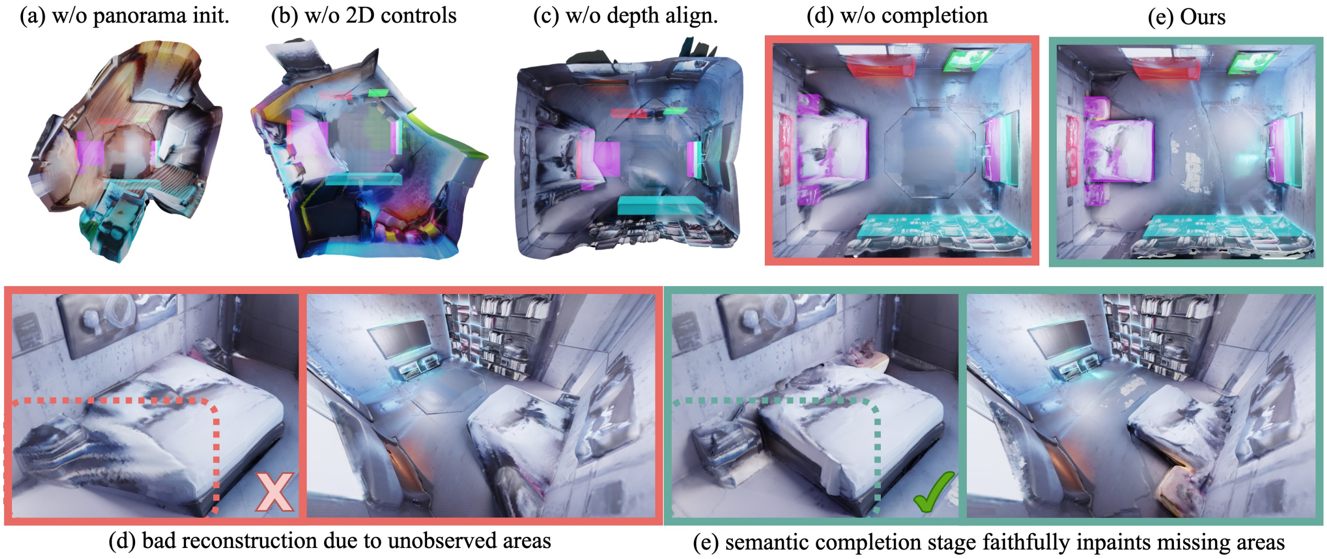
(very bad) ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 (very good)

[Perceptual Quality] On a scale of 1 to 5, how would you rate the scene overall? You might consider personal taste and how well the following prompt is captured in the scene: *

Rustic farmhouse style kitchen

(very bad) ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 (very good)

Figure 4. **User Study Interface.** 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).



(a) w/o panorama init. (b) w/o 2D controls (c) w/o depth align. (d) w/o completion (e) Ours

(d) bad reconstruction due to unobserved areas

(e) semantic completion stage faithfully inpaints missing areas

Figure 5. **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*).

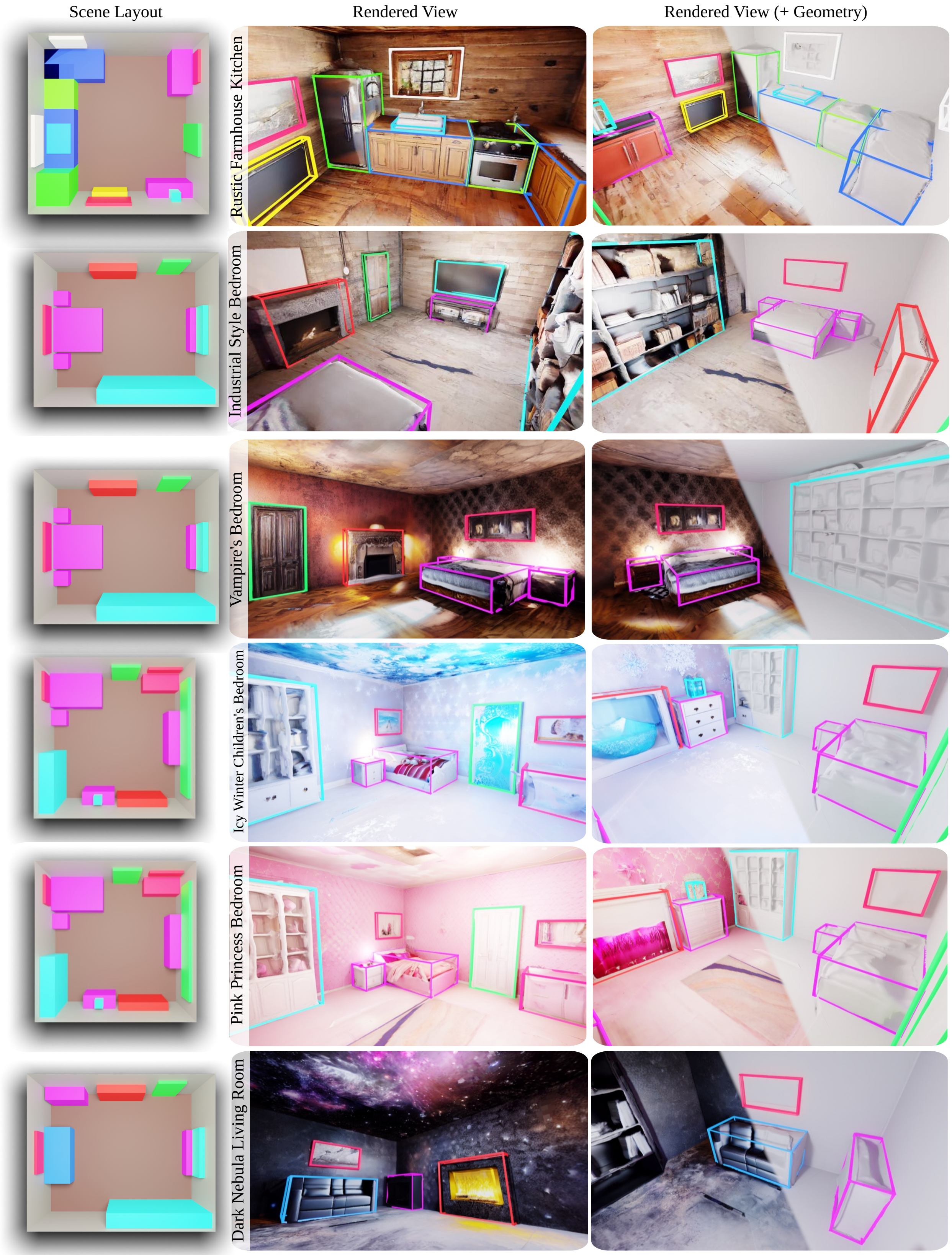


Figure 6. **Further Qualitative Examples.** We show colored geometry renderings of our method. ControlRoom3D generates convincing geometries and textures given a user-defined room layout and textual style description. More qualitative videos can be found at the end of the explanatory video in the supplementary material.