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, 18], posing a substantial barrier for end-users looking to create personalized scenes. Fueled by the recent progress in 2D generative models [13, 20, 40, 46], an emerging line of work aims to simplify the creation process of 3D objects [9, 27, 49, 53, 54] or even room-scale scenes [10, 16, 21, 58]. Among them, the recent work of Text2Room [21] creates compelling textured 3D room-scale meshes using a text prompt describing the scene and an off-the-shelf text-to-image model [40]. 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 [50] 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 [22, 32, 43–45, 52, 57] and powerful vision-language models [35], recent advancements in 2D diffusion models [5, 11, 31, 36, 40, 41] have significantly contributed to the ongoing progress in generating 3D objects and scenes. In particular, one prominent research direction focuses on single object synthesis [27, 30, 34, 49, 54, 58], wherein a differentiable 3D representation [23, 29] is commonly optimized through a loss signal derived from the denoising of rendered views [34, 53, 54]. Despite achieving impressive synthesis of object-centric scenes, the challenge of generalizing these techniques to large and complex scenes persists [54]. Typically, existing approaches rely on additional 3D data [3, 4, 12], which is challenging to obtain at scale. Related to our work, SceneScape [16] and Text2Room [21] use text-conditioned inpainting models [40] in combination with state-of-the-art mono-depth estimators [2, 37] to create a 3D mesh representation. Alternatively, LDM3D [48] directly predicts an RGBD frame based on a text prompt. However, these approaches have limited control over the generation process, typically leading to globally implausible room layouts. Concurrently, Fang *et al.* [15] thus focus on producing authentic room layouts through the generation of a global scene code from text, which is subsequently transformed into a 3D bounding box representation. In contrast, our work allows precise control over the scene composition by leveraging globally coherent 2D semantic and depth maps derived from 3D semantic proxy rooms. This approach not only guides but also aligns the generation process, ensuring that the scene geometry consistently corresponds with the proxy room's layout.

Semantically Guided 3D Scene Generation. An emerging line of works investigates scene generation with semantic compositing [8, 14, 17, 26, 40, 55, 56]. However, in 3D they are typically constrained to low resolution textures and partial room settings to enable inward-facing views [10, 33, 58] 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, 26, 55, 56, 59]. 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 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 [47, 50] or objects [28] 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.
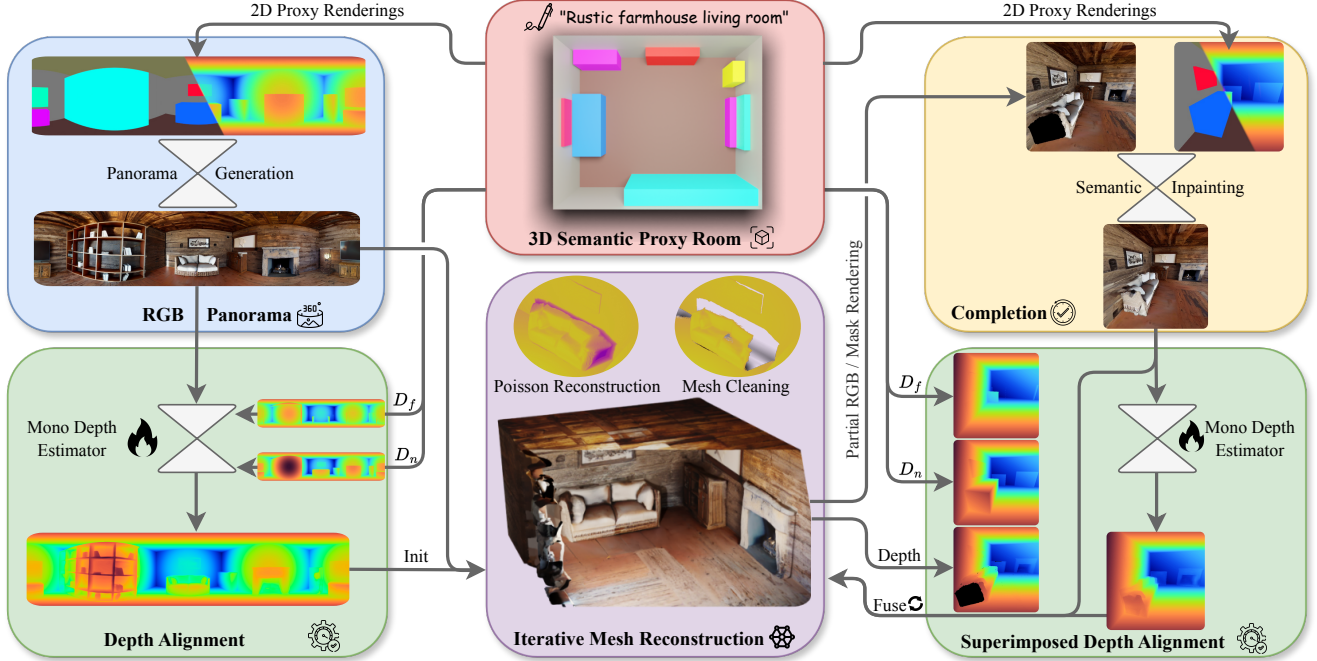
Figure 2. **Method Overview.** The *3D semantic proxy room* ▢ contains a user-defined 3D room layout based on semantic bounding boxes and a textual description of the overall room style. From this representation, we derive control signals that ensure 3D consistent textures and align corresponding depth 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 $\mathcal{M}_\circledS$ 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
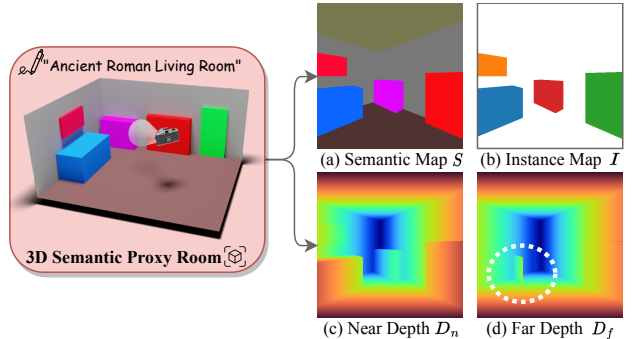

Figure 3. **Proxy Room Renderings.** We use classical rasterization to render semantic segmentation and instance maps as well as near and far depth maps from the proxy room.

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:

$$S, I, D_n, D_f = r(\mathcal{M}_\circledS, V)$$

Here, $r(\cdot)$ is standard rasterization without shading, the outputs $S \in [1, C]^{H \times W}$, $I \in \mathcal{N}^{H \times W}$, $D_n, D_f \in \mathbb{R}_+^{H \times W}$ are the rendered 2D semantic segmentation map of C classes, the instance segmentation map, and the near and far guidance
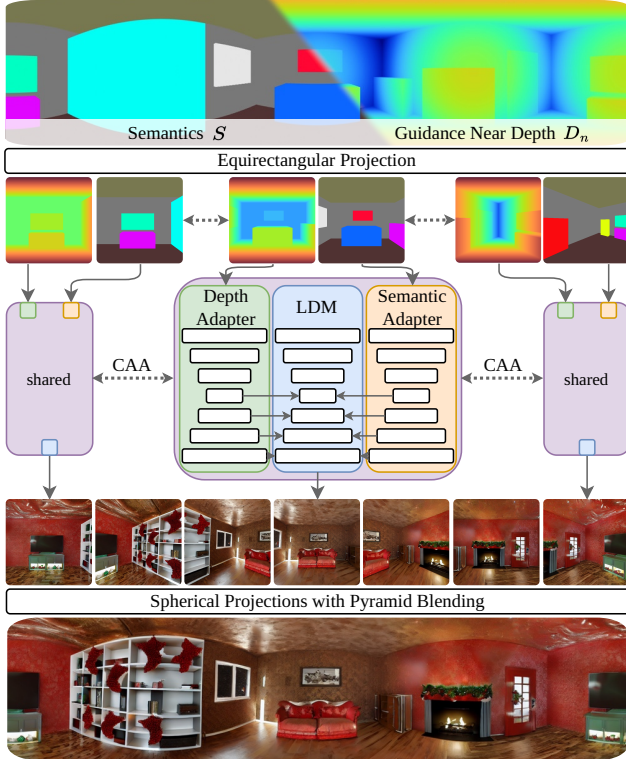
Figure 4. **Guided Panorama Generation.** We generate 360° panoramas guided by 3D semantic bounding boxes rendered from the proxy room. To this end, we equip an LDM □ with correspondence-aware attention modules (CAA) [50] and semantic □ and depth □ adapters finetuned on semantic bounding boxes.

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 observe that naively generating 2D views *individually* or only relying on iterative inpainting is detrimental for scene style consistency as the overall style tends to drift with each generated image and inaccurate loop closure [50]. Instead, we jointly create a comprehensive 360° view of the room that establishes a coherent and seamless style. This process is guided by the 3D semantic bounding boxes projected from the proxy room (Fig. 3a,c). To harness readily available powerful 2D generation models trained on perspective images, we first divide the semantic panorama into 8 views, each encompassing a 90° FOV and a 45° horizontal rotation using equirectangular projections, providing pixel-perfect correspondences between adjacent frames defined by 3×3 homography matrices. We then equip a latent diffusion model (LDM) with 3 types of modules. Firstly, correspondence-aware attention modules [50] use homographic correspondences between adjacent frames. Secondly, a semantic adapter guides the generation towards the specified room layout, while thirdly, the depth adapter con-
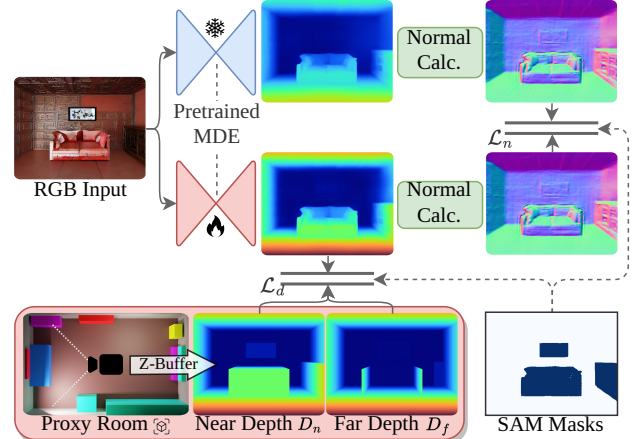

Figure 5. **Geometry Alignment.** We align generated 2D textures with the proxy room □. To achieve this, we derive both near and far depth maps by rendering from an identically posed camera inside the proxy room and align the generated objects (extracted with SAM [24]) accordingly by fine-tuning a pre-trained depth estimation network. Additionally, we preserve the overall object shape by ensuring the integrity of the normals during optimization.

ditions the generation on the guidance depth map. We employ pyramid blending to create the final panoramic image from the spherical projections of the generated images [50].

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 [39], 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.

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:

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

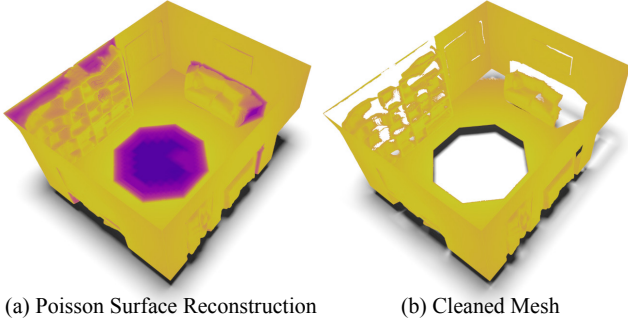(a) Poisson Surface Reconstruction   (b) Cleaned Mesh

Figure 6. **Mesh Cleaning.** After obtaining a watertight mesh with Poisson surface reconstruction (*left*), we delete vertices whose densities fall below a threshold (density is depicted with darker color shades). After removing small connected components, we obtain a cleaned mesh ready to be inpainted with novel content from favourable viewpoints (*right*).

Figure 7. **Completion Pipeline.** After inpainting small holes using classical Telea inpainting, we employ a 2D LDM □ guided by a semantic □ and depth □ adapter for larger missing areas, effectively inpainting these areas while aligning them to the control signals from the proxy room $\mathcal{M}_{\circledcirc}$.

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 [24] 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, □ 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
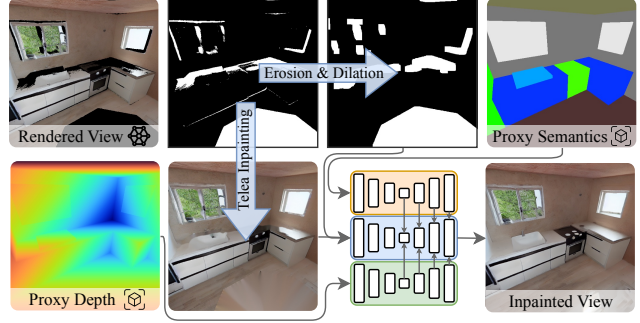
new content for these regions from a more favorable viewing angle. To this end, we first perform Poisson surface reconstruction, converting the 3D point cloud obtained from the RGBD panorama into a watertight mesh representation $\mathcal{M}_{\circledcirc}$ (Fig. 6a). We proceed to remove mesh vertices and their adjacent faces if their density falls below a specified threshold δ_v. Finally, we eliminate floating artifacts that are disconnected from the main mesh by deleting components with a vertex count below a specified threshold δ_c (Fig. 6b).

Mesh Completion. (Fig. 2, □ 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 $\mathcal{M}_{\circledcirc}$ with classical rasterization $r(\cdot)$ to obtain the RGB frame F, depth map $D_{\circledcirc}$ as well as the image mask M that identifies pixels without observed content: $F, D_{\circledcirc}, M = r(\mathcal{M}_{\circledcirc}, V)$. 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 [16, 21, 51]. 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, □). For seamless integration with pre-existing structures, we superimpose the rendered depth $D_{\circledcirc}$ 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.*, $D_n^{i,j} = D_f^{i,j} = D_{\circledcirc}^{i,j}$ for $M^{i,j} = 0$. 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) [19] and Inception score (IS) [42] on 100 RGB renderings from random viewpoints for each scene.

Implementation Details. We use a pre-trained LDM [40] as well as semantic and depth adapter modules. We finetune both the adapter weights on our 3D bounding box dataset derived from the HyperSim [39] dataset using BLIPv2 captions [25]. 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 [38] 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 [21] 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 [50] 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

Table 1. **Quantitative Comparison.** We report 2D metrics and user study results, including: CLIP Score (CS), Inception Score (IS), 3D Structure Completeness (3DS), Layout Plausibility (LP) and Perceptual Quality (PQ).

| Method	2D Metrics		User Study (3D Mesh)		
	CS ↑	IS ↑	3DS ↑	LP ↑	PQ ↑
Text2Room [21]	24.0	**6.47**	1.71	2.12	2.33
MVDiffusion [50]	25.6	6.37	2.48	3.01	2.89
ControlRoom3D (**Ours**)	**28.4**	5.75	**4.19**	**4.54**	**4.07**

Table 2. **Ablation Study.** (Fig. 2) We ablate the semantic proxy room (PR), geometry alignment module (GA), panorama generation (PG) and mesh completion (MC).

PR	GA	PG	MC	2D Metrics		User Study (3D Mesh)		
				CS ↑	IS ↑	3DS ↑	PA ↑	PQ ↑
✗	✗	✗	✗	27.0	5.76	1.78	1.14	2.33
✓	✗	✗	✗	27.6	5.60	2.81	2.02	2.95
✓	✓	✗	✗	27.8	5.64	3.34	4.34	3.45
✓	✓	✓	✗	**28.4**	**5.90**	3.68	4.55	3.95
✓	✓	✓	✓	**28.4**	5.75	**4.22**	**4.76**	**4.23**

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 provided on our project page.

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.
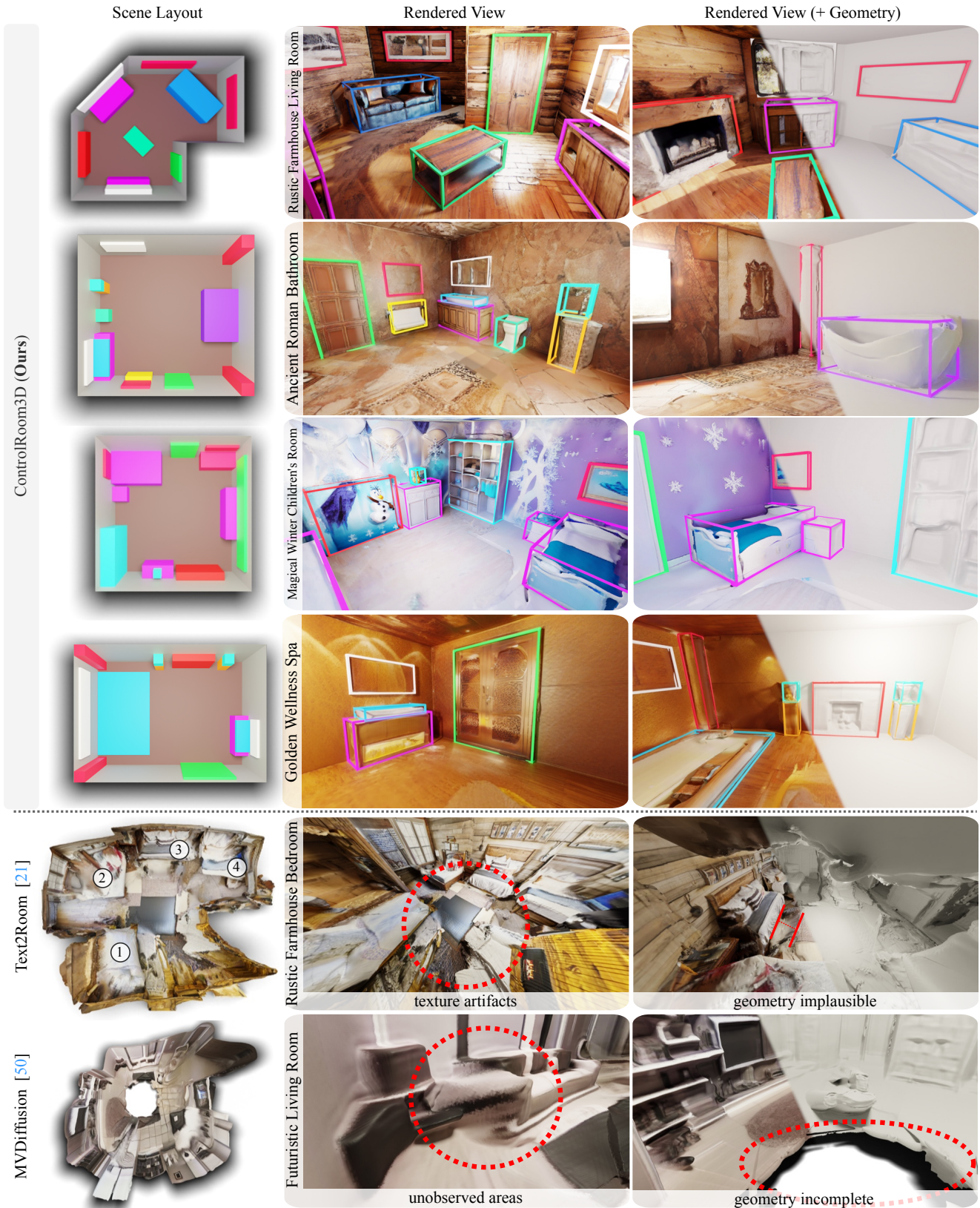
Figure 8. **Qualitative Comparison with Baseline Methods.** We show colored geometry renderings of our method and two baseline methods. ControlRoom3D generates convincing geometries and textures given a user-defined room layout and a textual style description. We provide further example images and videos on our project page.

Figure 9. **2D Guided Panorama Initialization.** We find that only using text prompts to generate panoramas result in subpar layouts, particularly, iterative inpainting repeats the most prominent object in the scene (*top, left*). Adding 2D guidance with semantic and depth maps, results in more plausible room layouts, whereas panorama generation yields a better style consistency (*bottom, right*).

Consistent Style Panorama Generation. The participants in our user study expressed a strong preference for the panorama generation module □, 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 [50] (Fig. 9a, *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. 9b, *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. 10, 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.

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).
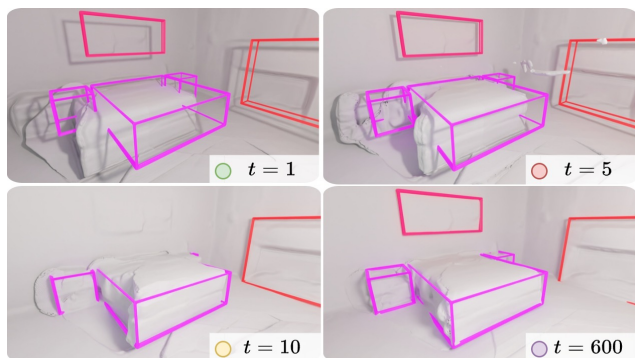

Figure 10. **Geometry Alignment.** Scale ambiguity leads to significant inaccuracies in MDEs (◯). In contrast, our proposed depth alignment module optimizes the alignment loss $\mathcal{L}$ to achieve strong alignment with the proxy room (◯).

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.

References

[1] Autodesk, Inc. Maya, 2019. 1

[2] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Iron-Depth: Iterative Refinement of Single-View Depth using Surface Normal and its Uncertainty. In *BMVC*, 2022. 2, 6

[3] Sherwin Bahmani, Jeong Joon Park, Despoina Paschalidou, Xingguang Yan, Gordon Wetzstein, Leonidas Guibas, and Andrea Tagliasacchi. Cc3d: Layout-conditioned generation of compositional 3d scenes. In *ICCV*, 2023. 2

[4] Miguel Angel Bautista, Pengsheng Guo, Samira Abnar, Walter Talbott, Alexander Toshev, Zhuoyuan Chen, Laurent Dinh, Shuangfei Zhai, Hanlin Goh, Daniel Ulbricht, Afshin Dehghan, and Josh Susskind. GAUDI: A Neural Architect for Immersive 3D Scene Generation. In *NeurIPS*, 2022. 2

[5] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving Image Generation with Better Captions. *OpenAI*, 2023. 2

[6] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. ZoeDepth: Zero-shot Transfer by Combining Relative and Metric Depth. *arXiv:2302.12288*, 2023. 2, 4, 6, 8

[7] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D Data in Indoor Environments. In *3DV*, 2017. 6

[8] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-Free Layout Control with Cross-Attention Guidance. *arXiv:2304.03373*, 2023. 2

[9] Yiwen Chen, Chi Zhang, Xiaofeng Yang, Zhongang Cai, Gang Yu, Lei Yang, and Guosheng Lin. IT3D: Improved Text-to-3D Generation with Explicit View Synthesis. *arXiv:2308.11473*, 2023. 1

[10] Dana Cohen-Bar, Elad Richardson, Gal Metzer, Raja Giryes, and Daniel Cohen-Or. Set-the-Scene: Global-Local Training for Generating Controllable NeRF Scenes. In *ICCVW*, 2023. 1, 2

[11] 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:2309.15807*, 2023. 2

[12] Terrance DeVries, Miguel Angel Bautista, Nitish Srivastava, Graham W. Taylor, and Joshua M. Susskind. Unconstrained Scene Generation with Locally Conditioned Radiance Fields. In *ICCV*, 2021. 2

[13] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 1

[14] Patrick Esser, Robin Rombach, and Björn Ommer. Taming Transformers for High-Resolution Image Synthesis. In *CVPR*, 2021. 2

[15] Chuan Fang, Xiaotao Hu, Kunming Luo, and Ping Tan. Ctrl-Room: Controllable Text-to-3D Room Meshes Generation with Layout Constraints. *arXiv:2310.03602*, 2023. 2

[16] Rafail Fridman, Amit Abecasis, Yoni Kasten, and Tali Dekel. SceneScape: Text-Driven Consistent Scene Generation. *arXiv:2302.01133*, 2023. 1, 2, 5

[17] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *ECCV*, 2022. 2

[18] Roland Hess. *Blender Foundations: The Essential Guide to Learning Blender 2.6*. Focal Press, 2010. 1

[19] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 6

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *NeurIPS*, 2020. 1

[21] 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

[22] 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

[23] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. In *ACM Transactions on Graphics*, 2023. 2

[24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment Anything. *arXiv:2304.02643*, 2023. 4, 5

[25] Junnan 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

[26] 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

[27] 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*, 2023. 1, 2

[28] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2Mesh: Text-Driven Neural Stylization for Meshes. In *CVPR*, 2022. 2

[29] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*, 2020. 2

[30] Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulo, Peter Kontschieder, and Matthias Nießner. Diffrf: Rendering-guided 3d radiance field diffusion. In *CVPR*, 2023. 2

[31] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and

Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2021. 2

[32] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. Combined scaling for zero-shot transfer learning. *Neurocomputing*, 2023. 2

[33] Ryan Po and Gordon Wetzstein. Compositional 3D Scene Generation using Locally Conditioned Diffusion. *arXiv:2303.12218*, 2023. 2

[34] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D Diffusion. In *ICLR*, 2022. 2

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2

[36] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv:2204.06125*, 2022. 2

[37] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer. *IEEE TPAMI*, 2020. 2

[38] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3D Deep Learning with PyTorch3D. *arXiv:2007.08501*, 2020. 6

[39] 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. 4, 6

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 6

[41] 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*, 2022. 2

[42] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved Techniques for Training GANs. In *NeurIPS*, 2016. 6

[43] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv:2111.02114*, 2021. 2

[44] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022.

[45] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *ACL*, 2018. 2

[46] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *ICML*, 2015. 1

[47] Liangchen Song, Liangliang Cao, Hongyu Xu, Kai Kang, Junsong Tang, Feng Yuan, and Yang Zhao. RoomDreamer: Text-Driven 3D Indoor Scene Synthesis with Coherent Geometry and Texture. *arXiv:2305.11337*, 2023. 2

[48] Gabriela Ben Melech Stan, Diana Wofk, Scottie Fox, Alex Redden, Will Saxton, Jean Yu, Estelle Aflalo, Shao-Yen Tseng, Fabio Nonato, Matthias Muller, and Vasudev Lal. LDM3D: Latent Diffusion Model for 3D. *arXiv:2305.10853*, 2023. 2

[49] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. DreamGaussian: Generative Gaussian Splatting for Efficient 3D Content Creation. *arXiv:2309.16653*, 2023. 1, 2

[50] 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, 4, 6, 8

[51] Alexandru Telea. An image inpainting technique based on the fast marching method. In *Journal of graphics tools*, 2004. 5

[52] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. In *Communications of the ACM*, 2016. 2

[53] 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*, 2023. 1, 2

[54] 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*, 2023. 1, 2

[55] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. BoxDiff: Text-to-Image Synthesis with Training-Free Box-Constrained Diffusion. In *ICCV*, 2023. 2

[56] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. ReCo: Region-Controlled Text-to-Image Generation. In *CVPR*, 2023. 2

[57] Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv:2309.02591*, 2023. 2

[58] Jingbo Zhang, Xiaoyu Li, Ziyu Wan, Can Wang, and Jing Liao. Text2NeRF: Text-Driven 3D Scene Generation with Neural Radiance Fields. *arXiv:2305.11588*, 2023. 1, 2

[59] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *ICCV*, 2023. 2