TEXT TO 3D OBJECT GENERATION FOR SCALABLE ROOM ASSEMBLY

Sonia Laguna *
ETH Zurich

Alberto Garcia-Garcia
Google

Marie-Julie Rakotosaona
Google

Stylianos Moschoglou
Google

Leonhard Helminger
Google

Sergio Orts-Escolano
Google

ABSTRACT

Modern machine learning models for scene understanding, such as depth estimation and object tracking, rely on large, high-quality datasets that mimic real-world deployment scenarios. To address data scarcity, we propose an end-to-end system for synthetic data generation for scalable, high-quality, and customizable 3D indoor scenes. By integrating and adapting text-to-image and multi-view diffusion models with Neural Radiance Field-based meshing, this system generates high-fidelity 3D object assets from text prompts and incorporates them into pre-defined floor plans using a rendering tool. By introducing novel loss functions and training strategies into existing methods, the system supports on-demand scene generation, aiming to alleviate the scarcity of current available data, generally manually crafted by artists. This system advances the role of synthetic data in addressing machine learning training limitations, enabling more robust and generalizable models for real-world applications.

1 INTRODUCTION AND BACKGROUND

Synthetic data has become a cornerstone in the training and development of machine learning (ML) models, particularly for perception tasks in computer vision. Applications spanning autonomous navigation, indoor scene understanding, and robotic interaction rely on vast amounts of high-quality, annotated training data that reflects real-world complexity (Nikolenko, 2021). However, acquiring such data at scale poses significant challenges, including privacy and fairness concerns, high annotation costs, and domain-specific constraints. Synthetic data generation offers an efficient and scalable solution, enabling the creation of diverse, controlled datasets that enhance model generalization to real-world environments. In the context of scene understanding, synthetic 3D indoor environments play a crucial role. Realistic room layouts populated with diverse and customizable assets provide valuable training grounds for perception tasks such as depth estimation, object segmentation, and scene reconstruction (Roberts et al., 2021; Zheng et al., 2020). By simulating variations in object type, color, placement, and material properties, synthetic data can bridge domain gaps that arise from the limited variability of real-world datasets. This ability to generalize across domains is pivotal for deploying perception models in unpredictable real-world scenarios.

Existing approaches to synthetic data generation and representation for 3D scenes span a wide range of methodologies, including procedural content creation (Raistrick et al., 2024), diffusion-based multi-view approaches (Gao et al., 2024), Neural Radiance Fields (NeRFs) (Mildenhall et al., 2021), 3D Gaussian Splatting (Kerbl et al., 2023), and 3D supervised latent generative models (Hong et al.). Meshes in 3D modeling are essential for bridging the mentioned abstract 3D representations with practical applications. By defining geometry through vertices, edges, and faces, they enable realistic rendering, efficient texture mapping, and seamless integration into graphics pipelines. Their lightweight structure ensures real-time performance, making them crucial for AR/VR, gaming, virtual prototyping, and scientific visualization. As demand for photorealistic 3D content grows, high-quality meshes remain vital for immersive and interactive experiences. This deepens the need for synthetic data in the form of 3D meshes. However, generating complete meshes of 3D scenes remains a challenge (Höllein et al., 2023). We include an extensive description of prior related work in Appendix A. In this work, by focusing on single asset generations in combination with existing synthetic scenes, we introduce a system to leverage prior synthetic data and scale up 3D indoor scene generations in a mesh format from a text prompt.

*Research conducted during an internship at Google. Correspondence to slaguna@inf.ethz.ch.

We introduce a hybrid system for scalable and customizable 3D room generation, combining pre-existing floor plans with prompt-engineered 3D asset creation. This approach enhances diversity, fidelity, and scalability while reducing reliance on manually designed datasets, supporting robust downstream ML models for perception tasks. By addressing common challenges, such as background speckles, global 3D consistency, and geometry fidelity, we incorporate refined prompt engineering, instance segmentation, object-based fine-tuning, and sparsity and normal regularization. Our key contributions are: (*i*) We propose an automated prompt generation system that scales synthetic asset creation while ensuring high diversity and consistency for indoor scenes. (*ii*) We extend and integrate state-of-the-art techniques in multi-view generation, NeRF representation, and meshing. Our improvements include fine-tuning diffusion models, applying sparsity loss functions, enhancing background segmentation, and leveraging advanced monocular depth estimation for precise surface normal supervision. (*iii*) We introduce an end-to-end system that integrates generated assets into artist-defined room layouts. We evaluate the synthesized environments showing their potential to bridge the gap between synthetic and real-world data, setting a new benchmark for synthetic scene generation in ML perception tasks. Figure 1 illustrates an overview of the proposed system.

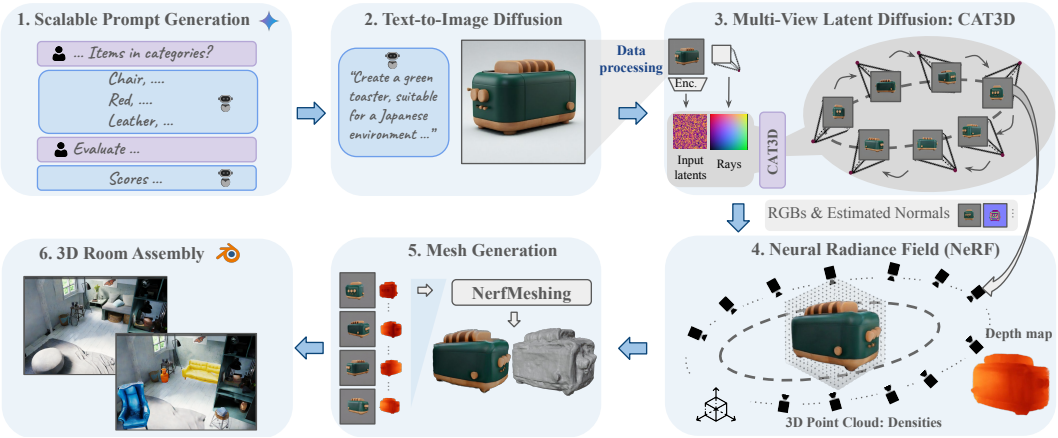

Figure 1: Overview of the proposed system from scalable prompt generation (**1**)-Section 2.1, through text-to-image diffusion (2)-Section 2.2, multi-view latent diffusion (3)-Section 2.3, and NeRF (4)-Section 2.4, onto the resulting meshes (**5**)-Section 2.5 integrated in existing rooms (**6**)-Section 2.6.

2 METHODS

The system integrates existing and adapted methods into a pipeline with subcomponents, as seen in Figure 1. This modular architecture grants flexibility to adapt to future advancements in ML research. The following section provides a detailed description of the components.

2.1 PROMPT ENGINEERING AND SCALABILITY

Prompt engineering plays a critical role in generating high-quality synthetic 3D scenes. Automated prompt generation expands object diversity in the synthetic data by incorporating domain-specific details. This involves designing precise textual inputs to ensure outputs meet desired specifications, including object type, material, color, and style. The nuances of prompt engineering significantly influence the quality, coherence, and scalability of generated assets.

Contextual Precision We conducted a study including specific contextual phrases to refine the text prompts. We found that adding in "in an empty white background" and "in the middle" improved spatial clarity, preventing issues like object cropping or awkward placements. These ensured that generated assets align with the pipeline's requirements, as shown in Figure 5 in Appendix B.

Scalable Prompt Generations We introduce a procedure for automating the generation of diverse, high-quality text prompts for synthetic room assets: (i) *Gemini* is used to generate category lists for objects, materials, colors, and high-level themes, each with 10 to 20 realistic entries, as described in

Steps 1 and 2 of Algorithm 1. For instance, objects might include chairs, tables, and vases, while materials range from wood and metal to fabric and glass. Descriptive colors such as blue or metallic gray further diversify the dataset. (ii) *Gemini* creates natural language templates (Step 3) to integrate these categories in various sentence structures, avoiding redundancy. Examples include: "Design a [Color] [Material] [Object] inspired by [High-level Theme] aesthetics," or "Create a [Object] that is both functional and aesthetically pleasing, using [Color] [Material] within a [High-level Theme] setting." (iii) Next, we combine the generated categories and templates to create diverse prompts by permuting the variables (Step 4). (iv) *Gemini* evaluates each prompt based on coherence, specificity, and creativity (Step 5), to ensure the prompts are realistic, plausible, and visually inspiring. (v) Finally, we rank the prompts based on the final score (Step 6) and save them as a *csv* file to feed them to the next step. This corresponds to Part 1 as shown in Figure 1. A comprehensive explanation of the procedure is included via pseudocode in Algorithm 1, and Table 1 shows examples of the different ranges of scores of the resulting prompts with each resulting generated mesh, both in Appendix B.

2.2 TEXT-TO-IMAGE GENERATION

This section builds upon previous prompt engineering by utilizing state-of-the-art text-to-image diffusion models, to generate high-quality 2D images for 3D reconstruction pipelines. Text prompts allow for scalable, adaptable, and diverse synthetic data generation, reducing reliance on real-world captures while offering fine-grained control over content (see part 2 in Figure 1). Furthermore, we integrate a segmentation step using DeepLab (Chen et al., 2017) to isolate objects and ensure smooth, homogeneous backgrounds, as well as adding a background with a padding of 20% around the object, key for isolated object generation. This design choice in particular addresses common issues in text-to-image models when the goal is to generate single objects. Generally, extraneous elements or cluttered backgrounds are present and can interfere. Segmentation improves the quality and consistency of inputs for 3D isolated object reconstruction. This refinement step is visualized as data processing between Part 2 and Part 3 in Figure 1.

2.3 MULTI-VIEW DIFFUSION

We build upon state-of-the-art multi-view latent diffusion model (LDM) approaches to generate globally consistent views. In particular, we focus on CAT3D (Gao et al., 2024), which has demonstrated superior performance in the literature. Prior models offering multi-view outputs like MV-Dream (Shi et al., 2023b), ImageDream (Wang & Shi, 2023), SyncDreamer (Liu et al., 2023a), SPAD (Cusini et al., 2022), SV3D (Voleti et al., 2024), and Zero123++ (Shi et al., 2023a) could be used instead to generate the image-to-3D assets. The CAT3D framework, in particular, extends traditional diffusion models by introducing a multi-view diffusion process that generates consistent 2D projections from single-image inputs and camera poses. Additionally, an autoregressive sampling of the views ensures geometric alignment across views, which is critical for accurate 3D reconstruction.

CAT3D addresses the challenge of generating coherent views from sparse inputs through the use of a multi-view latent diffusion model. The approach is structured around three core processes. First, latent representation encoding is performed by employing a variational auto-encoder (VAE) to transform input images into a latent space, capturing both spatial and semantic features necessary for reasoning about unseen views. Second, pose-conditioned diffusion ensures the model is conditioned on camera poses corresponding to observed and target viewpoints, achieved by concatenating a relative raymap representation of the camera's position and orientation with the latent embeddings. Third, progressive view synthesis is executed, where the model generates multiple target views by iteratively refining noisy latents into high-fidelity projections, thereby ensuring spatial consistency across synthesized outputs, as illustrated in Part 3 in Figure 1.

Consistency across synthesized views is paramount for downstream 3D reconstruction tasks. To this end, CAT3D incorporates innovations such as 3D self-attention layers and binary masks to indicate observed inputs during training, addressing the inherent challenges of sparse input scenarios. Strategic anchor views are selected using an autoregressive sampling strategy, enabling efficient reasoning from limited data. To scale to dense viewpoints, the model clusters proximal camera positions and leverages anchor views to maintain coherence across hundreds of target viewpoints. Robustness to sparse inputs is further enhanced by advanced loss functions, including LPIPS perceptual loss, which preserves semantic integrity even in conditions of high uncertainty. CAT3D generates all

kinds of 3D scenarios and multi-object scenes. However, full scenes pose a big challenge with the objective of generating 3D object mesh outputs. Therefore, in the system, we fine-tune CAT3D with internal high-quality single object datasets, specified in Section 3, to avoid background speckles or inconsistencies when shifting the focus from scene-based data.

2.4 NeRF to Create 3D Representations

NeRFs transform multi-view 2D projections into detailed 3D scene representations. They are neural network-based models that generate high-quality 3D scenes from 2D images by learning to represent volumetric density and view-dependent color at any 3D point, as illustrated in Part 4 in Figure 1. Our approach inherits ZipNeRF (Barron et al., 2023) with the modifications done for CAT3D and builds on it with novel enhancements, which allow for more robust and photorealistic 3D reconstruction focused on single objects, even from sparse data. These are: *Density Regularization* for better geometric fidelity to a single object, by including a logarithmic sparsity loss on the output densities (σ), $L_d = \text{mean}(\log(\sigma))$; *Input Segmentation* of all the input views using DeepLab for background consistency; *Normal Smoothness and Orientation Losses*, as introduced originally in Verbin et al. (2022); and *Estimated Normal Supervision*, by leveraging MariGold (MG) (Gupta et al., 2022), a state-of-the-art monocular depth estimation model. This last step uses the estimated depth to compute its gradients and use them as view-corresponding normals to regularize the geometry of the resulting asset. In particular, we compute the cosine loss on the estimated normal from the NeRF density, $\hat{\mathbf{n}}$, and the normal predicted from the MG approximation, $\hat{\mathbf{n}}^{\text{MG}}$. Note that by leveraging existing models we alleviate the need to train with geometry-informed 3D data. Figure 2 shows this step in a more comprehensive manner with the visualization of the predicted normals.

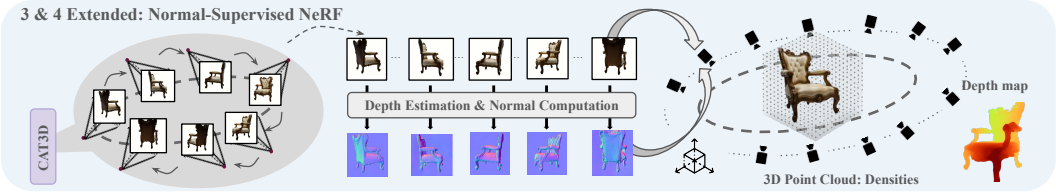

Figure 2: Normal estimation and its use in the NeRF, expanding on Steps 3 and 4 from Figure 1.

2.5 NeRFMeshing: Distilling NeRF into Meshes

NeRFMeshing (Rakotosaona et al., 2024) is a method designed to extract geometrically accurate 3D meshes from NeRFs, enabling their integration into standard computer graphics and simulation pipelines. By introducing a novel Signed Surface Approximation Network (SSAN), it distills the volumetric representation of NeRFs into a compact mesh format with precise geometry and view-dependent color properties. This process facilitates real-time rendering and supports physics-based simulations, overcoming traditional NeRF limitations in geometry accuracy and compatibility with graphics workflows. We use NeRFMeshing with additional (i) SSIM and (ii) LPIPS supervision for improved rendering after the mesh extraction.

2.6 Application: Placing Generated Objects in Rooms

We demonstrate the scalability of our system by incorporating generated 3D assets into complex room layouts using Blender, open source rendering software. The process includes: *Asset Scaling and Placement* by importing assets for intuitive scaling and alignment and *Scene Composition*, by positioning objects to achieve realistic spatial arrangements. This supports creating realistic, editable scenes for applications like simulation, architectural visualization, and interactive media.

3 Experimental Setup

Dataset: The data used to finetune CAT3D consists of internal high-quality datasets containing more than 300K isolated objects. These datasets highlight synthetic data's capacity to generalize

across domains. Furthermore, we apply the system to place objects in more than 64 rooms (crafted by 3D artists from Evermotion) in Blender.

Resources: The system uses 16 A100 GPUs to sample the multi-view diffusion model and to train and evaluate the NeRF and NerfMeshing architectures in up to 15 minutes. Note that prior work on mesh asset generation, i.e. TextMesh (Tsalicoglou et al., 2024), takes 2 hours to generate a single asset due to SDS optimization. Our system allows for flexibility of generations based on a prompt, as most steps have not been trained on specific 3D datasets, unlike prior models trained end-to-end with 3D-geometry-informed supervision that can struggle to generate out-of-domain when the input prompts are diverse. Our model is implemented in Jax v0.5 and room generation in Blender 4.3.

4 RESULTS

Our results highlight the effectiveness of our approach in 3D asset generation and scene composition, demonstrating geometric consistency, texture quality, and overall realism. Comprehensive visual evaluations establish our method as a new benchmark in the field.

4.1 EFFECT OF OBJECT SEGMENTATION, FINE-TUNING AND SPARSITY

To evaluate the effect of object segmentation, fine-tuning of CAT3D, and density regularization, we compare results generated by our system without these components against those generated with their inclusion. Without them, the generated assets lack precise boundaries and exhibit blending artifacts with the background, we see a significantly enhancement of the fidelity of textures and geometric details. Visual comparisons that illustrate the marked improvement in multi-view consistency and realism are displayed in Figure 3a.

4.2 NERF AND NERFMESHING GEOMETRY ADAPTATIONS

To evaluate the impact of our introduced modifications to NeRF, particularly the normal smoothness, orientation loss and normal supervision, we visualize assets generated from the same prompt with and without this components, as shown in Figure 3b. The introduced changes reduce artifacts and improve the quality of reconstructions, in particular visible with improved smoothness at the edges of the 3D meshes and in complex or sharp regions.

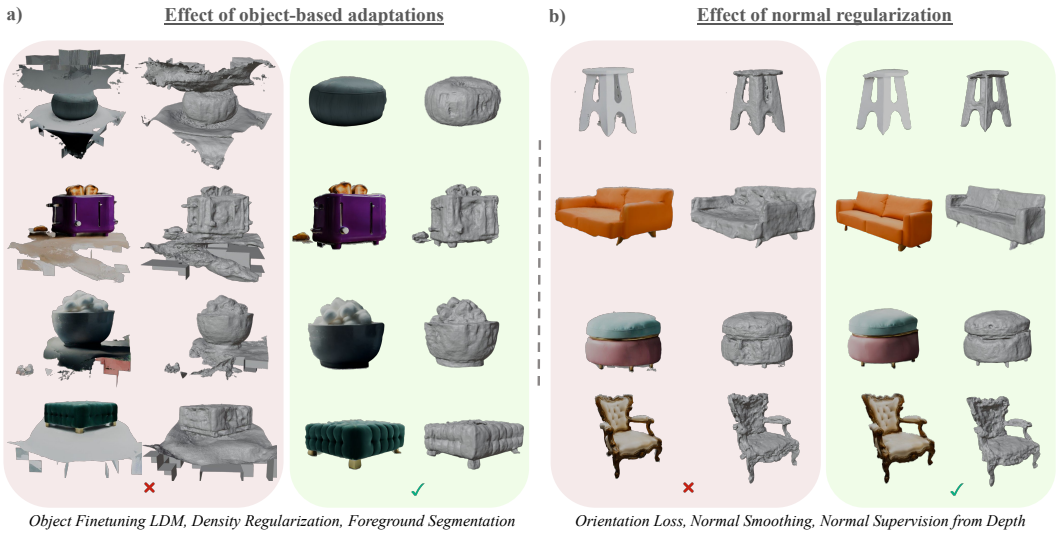

Figure 3: Qualitative evaluation of the contributions of our system to the different subcomponents. (a) Impact of object-based adaptations including segmentation, fine-tuning of the LDM, and density regularization. (b) Effect of NeRF and NeRFMeshing adaptations focusing on normal regularization. In particular, adding normal smoothness, orientation loss, and normal supervision.

4.3 QUALITATIVE RESULTS

To showcase the results of the complete end-to-end system, Figure 4 illustrates the results of Step 6 in Figure 1, the 3D Room Assembly. It displays an example starting from an original room (top left), where sequentially, annotated objects are replaced by semantic equivalents generated with the proposed system, to produce permutations of the same room. Further results for individually generated objects and failure modes have been included in Figure 6 in Appendix C.

Figure 4: Synthetic 3D room variations generated by the system. From the original room (top left), annotated objects are replaced by generated semantic equivalents to produce room permutations.

5 DISCUSSION AND CONCLUSION

We introduce an automated system that successfully generates high-fidelity 3D assets from text descriptions, leveraging state-of-the-art diffusion models, NeRF-based meshing, and advanced prompt engineering. The approach shows object consistency and realism while enabling scalable synthetic data generation for downstream ML and perception applications. Key contributions include fine-tuned multi-view diffusion, object segmentation, and normal estimation supervision, all of which significantly improve geometric accuracy and visual fidelity, as shown through visualizations. We show how these assets can be placed and ensure seamless integration into existing environments. Its relevance lies in addressing the key challenge of generating 3D models, which stems from the limited availability of 3D training data. This limitation results in significantly smaller datasets compared to those used in other domains like 2D image synthesis. Despite these advancements, challenges remain in handling complex multi-object scenes and automated object placement. Future work could focus on optimizing spatial arrangements. Our system controls synthetic data creation, bypassing privacy and copyright concerns, and sets a new benchmark in synthetic scene assembly and generation, bridging the gap between real and artificial data, with applications in virtual reality, gaming, and AI-driven simulation.

ACKNOWLEDGEMENTS

We thank the GLDM team at Google for their valuable contributions.

REFERENCES

Antonio Alliegro, Yawar Siddiqui, Tatiana Tommasi, and Matthias Nießner. Polydiff: Generating 3d polygonal meshes with diffusion models. *arXiv preprint arXiv:2312.11417*, 2023.

Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19697–19705, 2023.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848, 2017.

Zhiqin Chen, Andrea Tagliasacchi, and Hao Zhang. Bsp-net: Generating compact meshes via binary space partitioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 45–54, 2020.

Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4456–4465, 2023.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2920–2929, 2023.

Iris Cusini, Davide Berretta, Enrico Conca, Alfonso Incoronato, Francesca Madonini, Arianna Adelaide Maurina, Chiara Nonne, Simone Riccardo, and Federica Villa. Historical perspectives, state of art and research trends of spad arrays and their applications (part ii: Spad arrays). *Frontiers in Physics*, 10:906671, 2022.

Chuan Fang, Yuan Dong, Kunming Luo, Xiaotao Hu, Rakesh Shrestha, and Ping Tan. Ctrl-room: Controllable text-to-3d room meshes generation with layout constraints. *arXiv preprint arXiv:2310.03602*, 2023.

Rafail Fridman, Amit Abecasis, Yoni Kasten, and Tali Dekel. Scenescape: Text-driven consistent scene generation. *Advances in Neural Information Processing Systems*, 36, 2024.

Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Anchit Gupta, Wenhan Xiong, Yixin Nie, Ian Jones, and Barlas Oğuz. 3dgen: Triplane latent diffusion for textured mesh generation. *arXiv preprint arXiv:2303.05371*, 2023.

Youdh Chand Gupta, Sapna Panwar, Namita Banyal, Neelam Thakur, and MR Dhiman. Marigold. In *Floriculture and Ornamental Plants*, pp. 1–23. Springer, 2022.

Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7909–7920, 2023.

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. In *The Twelfth International Conference on Learning Representations*.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Jeong-gi Kwak, Erqun Dong, Yuhe Jin, Hanseok Ko, Shweta Mahajan, and Kwang Moo Yi. Vivid-1-to-3: Novel view synthesis with video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6775–6785, 2024.

Sixu Li, Chaojian Li, Wenbo Zhu, Boyang Yu, Yang Zhao, Cheng Wan, Haoran You, Huihong Shi, and Yingyan Lin. Instant-3d: Instant neural radiance field training towards on-device ar/vr 3d reconstruction. In *Proceedings of the 50th Annual International Symposium on Computer Architecture*, pp. 1–13, 2023.

Shichen Liu, Shunsuke Saito, Weikai Chen, and Hao Li. Learning to infer implicit surfaces without 3d supervision. *Advances in Neural Information Processing Systems*, 32, 2019.

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023a.

Zhen Liu, Yao Feng, Michael J Black, Derek Nowrouzezahrai, Liam Paull, and Weiyang Liu. Meshdiffusion: Score-based generative 3d mesh modeling. *arXiv preprint arXiv:2303.08133*, 2023b.

Weijia Mao, Yan-Pei Cao, Jia-Wei Liu, Zhongcong Xu, and Mike Zheng Shou. Showroom3d: Text to high-quality 3d room generation using 3d priors. *arXiv preprint arXiv:2312.13324*, 2023.

Luke Melas-Kyriazi, Christian Rupprecht, and Andrea Vedaldi. Pc2: Projection-conditioned point cloud diffusion for single-image 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12923–12932, 2023.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4460–4470, 2019.

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. *Communications of the ACM*, 65(1):99–106, 2021.

Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia 2022 conference papers*, pp. 1–8, 2022.

Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulo, Peter Kontschieder, and Matthias Nießner. Diffrf: Rendering-guided 3d radiance field diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4328–4338, 2023.

Sergey I Nikolenko. *Synthetic data for deep learning*, volume 174. Springer, 2021.

Goran Paulin and Marina Ivasic-Kos. Review and analysis of synthetic dataset generation methods and techniques for application in computer vision. *Artificial intelligence review*, 56(9):9221–9265, 2023.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, et al. Infinite photorealistic worlds using procedural generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12630–12641, 2023.

Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, et al. Infinigen indoors: Photorealistic indoor scenes using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21783–21794, 2024.

Marie-Julie Rakotosaona, Fabian Manhardt, Diego Martin Arroyo, Michael Niemeyer, Abhijit Kundu, and Federico Tombari. Nerfmeshing: Distilling neural radiance fields into geometrically-accurate 3d meshes. In *2024 International Conference on 3D Vision (3DV)*, pp. 1156–1165. IEEE, 2024.

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 *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10912–10922, 2021.

Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023a.

Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023b.

Jiapeng Tang, Yinyu Nie, Lev Markhasin, Angela Dai, Justus Thies, and Matthias Nießner. Diffuscene: Denoising diffusion models for generative indoor scene synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20507–20518, 2024.

Christina Tsalicoglou, Fabian Manhardt, Alessio Tonioni, Michael Niemeyer, and Federico Tombari. Textmesh: Generation of realistic 3d meshes from text prompts. In *2024 International Conference on 3D Vision (3DV)*, pp. 1554–1563. IEEE, 2024.

Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5481–5490. IEEE, 2022.

Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pp. 439–457. Springer, 2024.

Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 52–67, 2018.

Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023.

Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. *arXiv preprint arXiv:2311.12024*, 2023.

Zijie Wu, Yaonan Wang, Mingtao Feng, He Xie, and Ajmal Mian. Sketch and text guided diffusion model for colored point cloud generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8929–8939, 2023.

Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024a.

Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. In *European Conference on Computer Vision*, pp. 1–20. Springer, 2024b.

Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–16, 2023.

Jingbo Zhang, Xiaoyu Li, Ziyu Wan, Can Wang, and Jing Liao. Text2nerf: Text-driven 3d scene generation with neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 2024.

Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pp. 519–535. Springer, 2020.

Xin-Yang Zheng, Hao Pan, Peng-Shuai Wang, Xin Tong, Yang Liu, and Heung-Yeung Shum. Locally attentional sdf diffusion for controllable 3d shape generation. *ACM Transactions on Graphics (ToG)*, 42(4):1–13, 2023.

Xin-Yang Zheng, Hao Pan, Yu-Xiao Guo, Xin Tong, and Yang Liu. Mvdˆ 2: Efficient multiview 3d reconstruction for multiview diffusion. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024.

A RELATED WORK

3D Generation Techniques Recent advancements in neural generative models have significantly improved the quality and scalability of 3D object generation from single-view reconstruction tasks (Chen et al., 2020; Mescheder et al., 2019; Liu et al., 2019; Wang et al., 2018) to image-conditioned generative models (Melas-Kyriazi et al., 2023; Wu et al., 2023; Alliegro et al., 2023; Liu et al., 2023b; Cheng et al., 2023; Zheng et al., 2023; Gupta et al., 2023; Müller et al., 2023; Zhang et al., 2023). More recently, methods like Zero123++ (Shi et al., 2023a), ImageDream (Wang & Shi, 2023), MVDream (Shi et al., 2023b), SyncDreamer (Liu et al., 2023a) and CAT3D (Gao et al., 2024) leverage multi-view diffusion models to synthesize novel views with high cross-view consistency. By leveraging cross-view attention mechanisms, these approaches enhance 3D reconstruction from sparse input images, resolving depth ambiguities and ensuring globally coherent asset generation. Video-based diffusion models, like ViVid-1-to-3 (Kwak et al., 2024) and AnimateDiff (Guo et al., 2023), have also demonstrated the capability to simulate smooth camera trajectories and produce 3D-consistent representations, although with limited flexibility for static 3D object generation.

Similarly, Large Reconstruction Models (Hong et al.; Wang et al., 2023) map image tokens to implicit 3D representations such as triplanes or meshes through a transformer using multi-view supervision. By directly optimizing on mesh representations, approaches like InstantMesh (Xu et al., 2024a) and Instant3D (Li et al., 2023) integrate differentiable isosurface extraction modules to enhance geometric accuracy and reduce memory overhead, achieving efficient and high-quality 3D mesh generation from single images. These frameworks build on scalable transformer architectures, enabling generalization across diverse datasets. Extensions like GRM (Xu et al., 2024b) and MVD2 (Zheng et al., 2024) explore Gaussian-based representations and mesh-specific optimizations to improve rendering efficiency and surface quality. Despite their advancements, they often struggle with capturing fine details in complex scenes and require carefully curated 3D datasets for robust performance and geometry supervision, unlike the prior multi-view diffusion based approaches. This constrains them to the data distribution they are trained on, unlike the work proposed in this manuscript, where assets can be generated with total freedom of prompts.

Finally, procedural generation techniques, such as those employed in Infinite Photorealistic Worlds (Raistrick et al., 2023) and Infinigen Indoors (Raistrick et al., 2024), utilize algorithmic rules to create diverse and scalable datasets. These systems employ probabilistic programs and node-based tools like Blender to generate photorealistic indoor scenes. While highly effective for creating diverse training datasets, procedural approaches are often constrained by predefined content types and can be computationally expensive to enforce certain constraints, such as semantically rich asset integration in indoor environments.

NeRF-based Representations Neural Radiance Fields (NeRFs) have emerged as a powerful framework for 3D scene representation due to their ability to synthesize novel views with high fidelity (Mildenhall et al., 2021). Extensions such as Zip-NeRF (Barron et al., 2023) introduce regularization techniques to improve density, sparsity, and texture consistency, facilitating more efficient rendering and reducing artifacts. Advanced methods like Text2NeRF (Zhang et al., 2024) and Set-the-Scene (Cohen-Bar et al., 2023) integrate NeRFs with pre-trained text-to-image diffusion models, enabling text-driven 3D object and scene generation. Despite these advancements, NeRFs are inherently volumetric, making direct integration into standard graphics pipelines challenging. Approaches like NerfMeshing (Rakotosaona et al., 2024) address this limitation by converting radiance fields into explicit 3D meshes introducing a Signed Surface Approximation Network (SSAN) and using techniques such as isosurface extraction.

3D Mesh Generation Focusing on mesh-based representations, TextMesh (Tsalicoglou et al., 2024) extends NeRFs to Signed Distance Functions (SDFs) for precise mesh extraction, addressing issues such as oversaturation and texture inconsistency present in earlier methods like DreamFusion (Poole et al., 2022). This comes at the expense of a computational overhead by leveraging multi-view supervision and SDF backbones, where each asset takes around 2 hours to generate. InstantMesh (Xu et al., 2024a), on the other hand, combines multi-view diffusion models with sparse-view reconstruction frameworks to generate 3D meshes directly from single images. This approach integrates geometric supervision on the mesh surface using differentiable isosurface extraction modules. Despite these advancements, existing mesh generation approaches still struggle

with producing highly detailed meshes in complex scenes and maintaining consistent performance across diverse object categories.

Text-to-3D and Synthetic Data for Scene Understanding Recent advancements in text-driven 3D generation have evolved from early CLIP-guided methods, such as CLIPMesh (Mohammad Khalid et al., 2022), to more sophisticated diffusion-based systems leveraging large-scale text-to-image models (Zhang et al., 2024). Prompt engineering is critical to improving object placement, spatial alignment, and scene composition. Techniques like ShowRoom3D (Mao et al., 2023) and SceneScape (Fridman et al., 2024) refine 3D scene synthesis using score distillation and camera trajectory optimization, but often require extensive fine-tuning and are computationally intensive.

Synthetic data generation helps address the challenge of acquiring annotated datasets for scene understanding (Paulin & Ivasic-Kos, 2023), with methods like Text2Room (Höllein et al., 2023), SceneScape (Fridman et al., 2024), and Ctrl-Room (Fang et al., 2023) using text-to-image models to generate 3D meshes, though they face challenges with geometric distortion, occlusion handling, and scalability. Conditional diffusion models, such as DiffuScene (Tang et al., 2024), enhance semantic control over scene composition. Despite these improvements, many methods still struggle with achieving global consistency, geometry fidelity, and artifact-free outputs, especially in complex or occluded scenes. A limitation in these approaches is the dependence on additional 3D supervision, such as geometry through depth or normal maps. In contrast, our method leverages a multi-view Latent Diffusion Model (LDM) pre-trained on RGB instances and camera poses, eliminating the need for extra 3D supervision while maintaining high-quality mesh generation. Additionally, we are not restricted by the domain of the data on which the mesh generation was trained, unlike existing approaches. This grants us the capability to generate assets from previously unseen distributions and enhances the scalability of the later-assembled synthetic rooms. By focusing on generating isolated assets and placing them in existing 3D rooms, we scale existing room data and improve the quality and scalability of synthetic 3D indoor scene generation, bridging the gap between synthetic and real-world data.

B PROMPT ENGINEERING RESULTS

Contextual Precision Figure 5 shows examples of the effect of the contextual precision in the generated outputs, as discussed in Section 2.1. By adding in "in an empty white background" and "in the middle", we improve spatial clarity (Right), preventing issues like object cropping or awkward placements (Left).

Figure 5: Samples of the effect of the contextual precision in the generated outputs.

Scalable Prompt Generations A comprehensive explanation of the procedure to generate prompts in a scalable manner is included via a pseudocode in Algorithm 1, as explained in Section 2.1. Table 1 shows examples of the different ranges of scores of the resulting prompts together with the corresponding generated asset.

Algorithm 1 Structured Dataset Creation and Evaluation for 3D Object Text Prompts

Input: Initial category lists: {*Objects, Materials, Colors, High-level Themes*}
Output: Final dataset of ranked and filtered text prompts ready for 3D asset generation.

Step 1: Generate Initial Category Lists
Prompt LLM to create 10 to 20 diverse and realistic items for each category.
Organize results into a structured table.

Step 2: Refine and Validate Category Lists
Remove duplicates and nonsensical combinations from raw lists.

Step 3: Generate Structured Prompt Templates
Define template variables: {*Object, Color, Material, High-level Theme*}.
Prompt LLM to generate 20 templates with diverse sentence structures.

Step 4: Combine Categories to Create Text Prompts
Randomly sample combinations of {*Object, Material, Color, High-level Theme*}.
Fill placeholders in structured templates with sampled values.

Step 5: Evaluate and Refine Prompts
Feed generated prompts into LLM for evaluation based on:
(a) **Coherence:** Does the object-material-color combination make sense?
(b) **Specificity:** Is the object and setting clearly described?
(c) **Creativity:** Does the prompt inspire visually compelling designs?
Assign a realism score (1-10) to each prompt, with explanations.

Step 6: Rank and Output Results
Sort prompts by realism scores and output resulting list as *csv* file.

return Final dataset of filtered and ranked prompts.

Table 1: Samples of prompts, scores, and corresponding figures resulting from the proposed prompt engineering approach. The explanations behind the scores given by the *Gemini* are: *Coherence* - Materials and colors are plausible for the object, *Specificity* - Clarity of the design details, and *Creativity* - Suggests visually compelling design elements.

Prompt	**Score**	**Mesh**
Create a green toaster, suitable for a Japanese environment in an empty white background and in the middle.	**9.0**: Coherence 9/10, Specificity 9/10, Creativity 9/10.	
An antique concept: a white armchair that seamlessly integrates into the design in an empty white background and in the middle.	**8.0**: Coherence 9/10, Specificity 7/10, Creativity 8/10.	
An industrial interpretation of a plant, realized in pink ceramic in an empty white background and in the middle.	**7.0**: Coherence 5/10, Specificity 9/10, Creativity 7/10.	
Design a vase that is black and crafted from cotton reflecting the core principles of Art Deco design in an empty white background and in the middle.	**5.4**: Coherence 3/10, Specificity 9/10, Creativity 4/10.	

C ADDITIONAL QUALITATIVE RESULTS

Figure 6: Visualization of generated 3D objects. Note samples of failure cases where holes are present due to reconstruction errors.