GaussianTeller: Native 3D Gaussian Generation

Başak Melis Öcal*
University of Amsterdam

Xiaoyan Xing
University of Amsterdam

Ngo Anh Vien
BCAI, Robert Bosch GmbH

Sezer Karaoğlu
University of Amsterdam

Theo Gevers
University of Amsterdam

Abstract

3D content generation is important for fields like gaming and VR, requiring scalable methods that produce high-fidelity assets with real-time rendering capabilities — making 3D Gaussian splats a promising and efficient representation. While text/image-to-3D has advanced across various representations, native Gaussian splat generation remains underexplored due to their discrete, unstructured nature. To address this, we introduce GaussianTeller, a novel framework that learns robust diffusion priors from spatially-grouped Gaussians encoded into a structured latent space. A latent diffusion model denoises noisy latent codes conditioned on text or image inputs, while disentangled geometry and appearance decoders reconstruct high-fidelity Gaussian parameters. Our native design enables scalable, generalizable 3D generation and unlocks downstream tasks such as feed-forward Gaussian splat editing.

1. Introduction

3D content generation is becoming a core requirement in fields such as game development, virtual reality, and digital arts, where immersive and controllable 3D experiences are increasingly in demand. Meeting the requirements calls for scalable generation methods that produce high-fidelity 3D assets with high-quality rendering capabilities.

Inspired by recent breakthroughs in text-to-image diffusion models [3, 22], DreamFusion [18] pioneered to use pretrained diffusion models to optimize 3D representations via Score Distillation Sampling (SDS). Follow-up work improved fidelity [10, 26] and accelerated optimization by combining 3D Gaussian Splatting (3DGS) [7] with 2D diffusion models [30]. However, SDS-based techniques suffer from slow per-scene optimization and geometric ambiguities. Reconstruction-based methods [5] offer feed-forward alternatives, enabling direct 3D generation from single-view images. Nonetheless, these deterministic approaches struggle to represent uncertainties intrinsic to single-view 3D reconstruction. To address this, recent works fine-tune 2D diffusion models for multi-view image generation [12], extended further using video diffusion models [25]. Building on LRM, two-stage pipelines like Instant3D [9] synthesizes multi-view, then reconstructs 3D representations. Follow-up works improve geometry via differentiable mesh extraction [28] or regress 3D Gaussians from pixels or features [24]. More recent techniques [11] perform latent space diffusion on encoded multi-view splatter images. However, these solutions still fundamentally depend on multi-view inputs, lacking a true 3D latent space, limiting geometric precision.

Recent works have extended diffusion models to 3D representations [13, 15, 20, 23]. Approaches like GaussianCube [32] and GVGEN [2] adapt diffusion to 3D Gaussians via volumetric structures, but introduce costly preprocessing and training. To improve scalability, latent diffusion methods [4, 6, 29, 31] have been introduced. Specifically, L3DG [21] and DiffGS [33] apply latent diffusion to 3D Gaussians more efficiently, but remain category-specific and struggle to generalize across diverse object types. Despite these advances, native 3D Gaussian generation remains underexplored due to the inherent complexity of modeling their unstructured and discrete nature.

To address the limitations of previous works, we introduce *GaussianTeller*, a novel native 3D Gaussian splat generation framework that directly learns robust 3D diffusion priors from spatially-grouped Gaussians encoded into a structured latent space. A Gaussian latent diffusion model then learns denoising the noisy latent conditioned on text or image. Geometry and color decoders, together with a cross-branch feature sharing module, predict features conditioned

[*] Correspondence to b.m.ocal@uva.nl

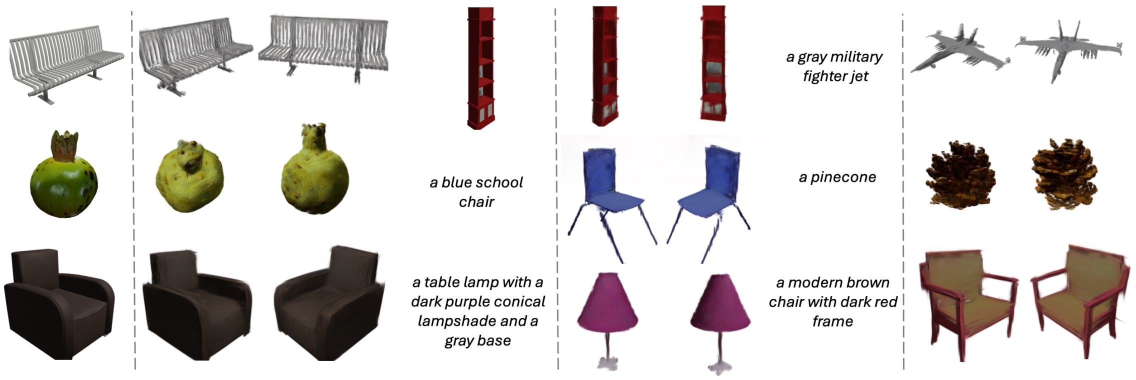
a gray military fighter jet
a blue school chair
a pinecone
a table lamp with a dark purple conical lampshade and a gray base
a modern brown chair with dark red frame

Figure 1. We present *GaussianTeller*, a native 3D Gaussian splat generation framework that learns robust 3D diffusion priors from spatially-grouped Gaussians encoded into a structured latent space. *GaussianTeller* enables effective generation of high-quality, 3D-consistent assets.

on spatial anchors, then reconstruct the Gaussian parameters. This Gaussian group-structured latent representation effectively manages the discrete nature of Gaussian splats and ensures enhanced 3D consistency, while spatial anchor conditioning aids 3D geometry recovery. Since the framework directly operates on input Gaussians, it also enables downstream tasks such as feed-forward 3DGS editing, unlocking new possibilities for scalable and generalized 3D generative modeling. In brief, our contributions include:

- A native 3D Gaussian splat generation framework, producing high-quality, 3D consistent assets.
- An effective strategy for learning 3D diffusion priors within a structured latent space, along with a disentangled geometry and appearance prediction mechanism for 3DGS.
- To the best of our knowledge, we are the first to directly operate on 3D Gaussian splats in a generalizable manner beyond category-specific priors.

2. Method

We introduce *GaussianTeller*, a native 3D Gaussian splat (3DGS) diffusion model that learns 3D priors directly from Gaussian groups encoded into a latent space via a 3D Gaussian VAE. The training pipeline is divided into two stages:

1. An encoder $\mathcal{E}_\theta$ maps the grouped input Gaussians $\mathcal{G}$, into a 3D latent $\mathbf{z}$. Dual-decoders $\mathcal{D}_{g_\phi}$ and $\mathcal{D}_{c\psi}$, together with a cross-branch feature sharing module f_ϑ, first predict geometry and appearance features conditioned on spatial anchors, then reconstruct the Gaussian parameters to produce $\mathcal{G}'$.
2. A Gaussian latent diffusion model d_ϱ learns denoising the noisy latent $\mathbf{z}_T$ conditioned on embedding $\mathcal{C}$.

At inference, *GaussianTeller* generates high-quality, 3D-consistent assets from text or image. To enhance generalizability, it is trained on a mix of public datasets and curated samples. The scheme of the framework is illustrated in Fig. 2.

2.1. Gaussian VAE for Group-structured Latent Space Learning

The unstructured nature of 3DGS data hinders spatial structure recovery in native generation. *GaussianTeller* addresses this by grouping input Gaussians by spatial proximity and encoding them into a latent space that retains global 3D structure for effective diffusion prior learning.

Encoder. More formally, the input is a set of Gaussian splats $S = \{S_i\}_{i=1}^{N}$, where after stacking the attributes we have $S = [\boldsymbol{\mu}, \mathbf{r}, \mathbf{s}, \mathbf{c}, \mathbf{o}] \in \mathbb{R}^{N\times 14}$, with each splat parameterized by its centroid $\boldsymbol{\mu} \in \mathbb{R}^{N\times 3}$, scale $\mathbf{s} \in \mathbb{R}^{N\times 3}$, quaternion vector $\mathbf{r} \in \mathbb{R}^{N\times 4}$, opacity $\mathbf{o} \in \mathbb{R}^{N\times 1}$ and color $\mathbf{c} \in \mathbb{R}^{N\times 3}$, representing a 3D asset. First S is randomly downsampled to a fix number of splats $\hat{S} \in \mathbb{R}^{\hat{N}\times 14}$. We follow the encoding strategy of GaussianMAE [14], and group the Gaussians using their centroids, then compute group centers $\boldsymbol{\mu_g} = \text{FPS}(\boldsymbol{\mu}) \in \mathbb{R}^{p\times 3}$ via farthest point sampling, where p denotes the number of groups. We refer $\boldsymbol{\mu_g}$ as *spatial anchors*. For each anchor, k neighboring splats are then obtained as $G = \text{KNN}(\hat{S}, \boldsymbol{\mu_g}) \in \mathbb{R}^{p\times k\times 14}$. The resulting Gaussian groups G are then fed into a tokenizer to obtain group tokens, $T = \text{Tokenizer(G)}$, with $T \in \mathbb{R}^{1024\times 1024}$. Finally, the group tokens are processed by the transformer-based encoder $\mathcal{E}_\theta$ introduced in [14]:

$$\mathbf{z} = \mathcal{E}_\theta(T) \quad \text{and} \quad \mathbf{z} \in \mathbb{R}^{1024\times 1024}, \tag{1}$$

where $\mathbf{z}$ is the Gaussian group-structured latent representing the 3D input.

Decoder. Another challenge in 3D Gaussian generation is the joint learning of Gaussian parameters, which can exhibit an uneven distribution, as pointed out by [14]. To predict these parameters from $\mathbf{z}$, we employ a dual-decoder architecture with a lightweight cross-branch feature sharing module f_ϑ. The geometry decoder $\mathcal{D}_{g_\phi}$ predicts geome-

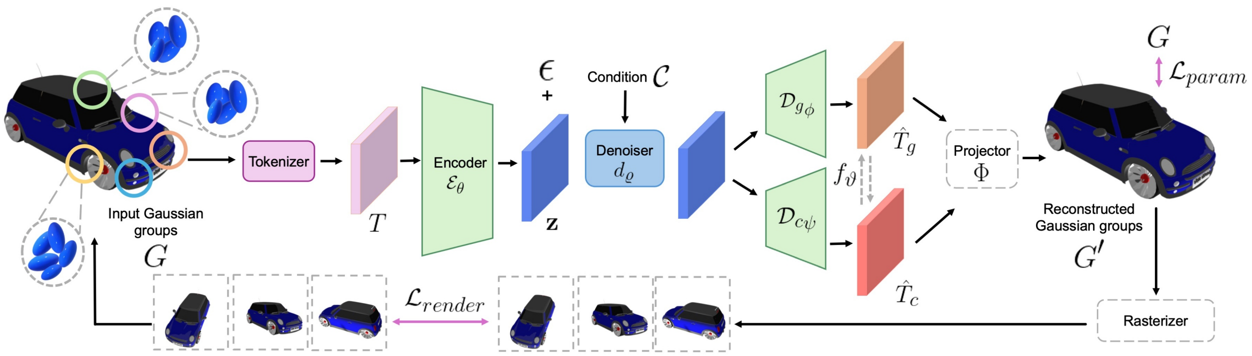

Figure 2. **Overview of our method. First stage:** Given input Gaussians, our method first groups them based on spatial proximity and encodes into a group-structured latent $\mathbf{z}$. Decoders $\mathcal{D}g\phi$ and $\mathcal{D}c\psi$ predict geometry and appearance from spatial anchors to reconstruct Gaussians $\mathcal{G}'$. **Second stage:** A denoiser d_ϱ learns to denoise $\mathbf{z}_T$, guided by condition $\mathcal{C}$.

try tokens T_g for centroids, scales and rotation quaternions, while the appearance decoder $\mathcal{D}_{c\psi}$ produces appearance tokens T_c for color and opacity. The transformer-based decoders are conditioned on positional embeddings $\varphi(\boldsymbol{\mu_g})$ derived from spatial anchors $\boldsymbol{\mu_g}$ to facilitate 3D structure recovery:

$$T_g = \mathcal{D}_{g_\phi}(\mathbf{z}, \varphi(\boldsymbol{\mu_g})), \quad T_c = \mathcal{D}_{c\psi}(\mathbf{z}, \varphi(\boldsymbol{\mu_g})) \tag{2}$$

These tokens are then passed to the cross-branch feature sharing module f_ϑ to yield a fused token $T_f = f_\vartheta(T_g, T_c)$, which is then simply integrated into the original tokens via a residual connection to enforce consistency between the geometry and appearance branches:

$$\hat{T}_g = T_g + T_s, \quad \hat{T}_c = T_c + T_s \tag{3}$$

Finally, a projection module Φ predicts the Gaussian attributes for the reconstructed groups G',

$$\boldsymbol{\mu}', \mathbf{r}', \mathbf{s}' = \Phi(T_g), \quad \mathbf{c}', \mathbf{o}' = \Phi(T_c) \tag{4}$$

Training. To capture 3D priors effectively, we supervise the Gaussian VAE with an $\mathcal{L}_1$ parameter reconstruction loss computed between the input and reconstructed groups:

$$\mathcal{L}_{param} = \|G, G'\|_1 \tag{5}$$

Due to the many-to-one mapping between the input 3D Gaussian parameters and images they are trained from, an $\mathcal{L}_1$ loss alone can be overly restrictive. Incorporating a render loss $\mathcal{L}_{render}$ relaxes this constraint and allows the model to capture a broader range of valid solutions. The render loss is computed over V random views of the images from which the input Gaussians are optimized, and is complemented by an LPIPS loss.

$$\mathcal{L}_{render} = \mathcal{L}_{rgb} + \beta\mathcal{L}_{LPIPS} \tag{6}$$

The Gaussian VAE is trained end-to-end with the following objective:

$$\mathcal{L}_{GVAE} = \lambda_1\mathcal{L}_{param} + \lambda_2\mathcal{L}_{render} + \lambda_3\mathcal{L}_{loc} + \lambda_4\mathcal{L}_{KL}, \tag{7}$$

where $\mathcal{L}_{KL}$ is the KL-divergence loss and $\mathcal{L}_{loc}$ is the locality-preserving regularization loss that encourages smooth color transitions within the Gaussian groups.

2.2. Gaussian Latent Diffusion

Our Gaussian VAE maps any 3DGS into a group-structured latent preserving global 3D structure. To train a diffusion model in this space, we employ the tranformer-based text/image-conditioned denoiser of Shap-e [6], and its pretrained weights for initialization as Shap-e is trained on 3D assets encoded into a latent space, and already carries rich 3d priors. Although the distribution learned by the Shap-e denoiser differs from that of our Gaussian group-structured latent space, this initialization provides a valuable starting point for further adaptation.

Given Gaussian group-structured latent $\mathbf{z} \in \mathbb{R}^{1024\times1024}$ encoded by $\mathcal{E}_\theta$ and a text/image condition $\mathcal{C}$ (obtained by encoding image/text via CLIP [19]), the forward process gradually adds Gaussian noise, corrupting $\mathbf{z}$ into $\mathbf{z_T}$ over T time steps. The denoiser d_ϱ is then tasked with recovering the latent distribution from $\mathbf{z_T}$. The training objective is:

$$\mathbb{E}_{\mathbf{z},t,\epsilon\sim\mathcal{N}(0,I)} \left\|d_\varrho(\mathbf{z_t}, \mathcal{C}, t) - \mathbf{z})\right\|^2, \tag{8}$$

where t denotes the time steps, and ϵ is noise level sampled from $\mathcal{N}(0, I)$. Classifier-free guidance is also employed.

2.3. 3D Asset Generation

Our method uses spatial anchors to condition the decoding process and preserve the spatial structure of 3D assets. At generation time, given a text or image condition, our framework first produces a coarse point cloud using an off-the-shelf model [16]. In practice, the point cloud can be any user-provided input and is used solely to initialize the spatial anchors $\boldsymbol{\mu_g}$ by the procedure in Sec. 2.1. We then sample the latent $\mathbf{z}$, and decode it, conditioned on the spatial anchors, to generate the final 3D asset. To ensure generalizability across both generated and user-provided point

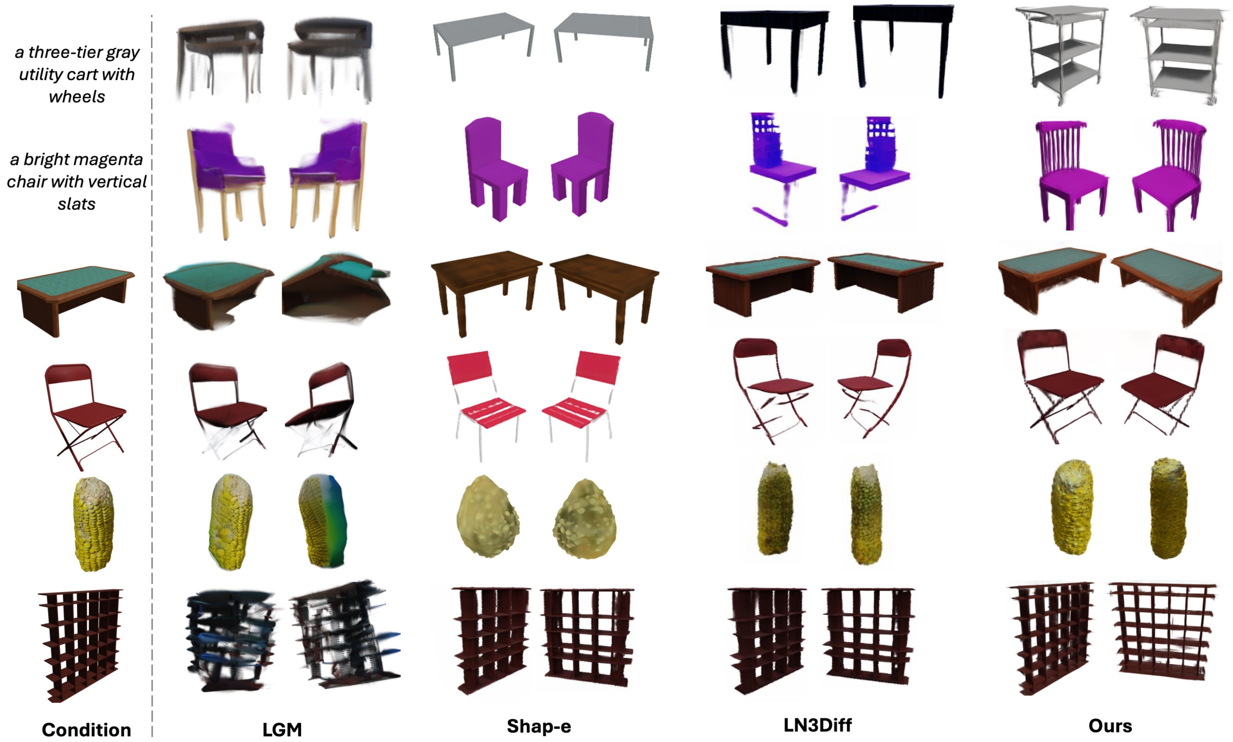
a three-tier gray utility cart with wheels
a bright magenta chair with vertical slats
Condition
LGM
Shap-e
LN3Diff
Ours

Figure 3. **Qualitative comparison with state-of-the-art methods on text/image conditioned 3D generation.** *GaussianTeller* produces high-quality, 3D-consistent assets that align with the given conditioning input.

clouds, our framework is trained on the dataset described in Sec. 2.4.

2.4. Dataset Assembly and Curation

To enable robust generalization in native 3D generation, we merge two complementary datasets and curate a third for training: 1) ShapeSplatsV1 [14] provides 52K 3DGS assets from 55 categories with 72 views each. 2) OmniObject3D [27] comprises 6K objects in 190 categories, with 100 random views. Their 3DGS representations are produced by LightGaussian [1]. 3) We curate 20K samples using LGM [24]. Curated assets are rendered from multiple viewpoints. DINOv2 [17] embeddings of rendered and dataset views are compared via cosine similarity, discarding assets with low similarity. A random view is selected from all available views of a 3D asset for image conditioning.

3. Results

To evaluate text/image conditioned generation, an unseen subset of our compiled dataset is employed. Our approach is evaluated against three state-of-the-art open-source methods for text/image conditioned 3D generation [6, 8, 24]. Fig. 3 presents a visual comparison of *GaussianTeller* for text/image conditioned 3D generation, while Fig. 1 shows additional qualitative results. Our method yields superior alignment with the input text, whereas other methods struggle to accurately follow the prompt and produce high-quality geometry. For image-to-3D generation, [24] produces inconsistent results due to the ambiguity of single-view reconstruction. Shap-E [6] shows misalignment with the input image, and LN3Diff [8] exhibits artifacts such as disconnected components and reduced geometric fidelity. In contrast, *GaussianTeller* generates high-quality, 3D-consistent outputs that closely align with the given image condition.

4. Conclusion

We introduced a native 3DGS generation framework that produces high-quality, 3D-consistent assets. By learning diffusion priors within a structured latent space, our method effectively models discrete Gaussians. Disentangling geometry prediction from appearance proves to improve control and accuracy in reconstructing complex 3D structures. As a native approach trained directly on 3D splats, *GaussianTeller* opens new directions for downstream tasks such as editing.

Acknowledgements

We thank members of the Bosch-UvA Delta Lab and anonymous reviewers for helpful discussions and feedback. This project was generously supported by the Bosch Center for Artificial Intelligence.

References

[1] Zhiwen Fan, Kevin Wang, Kairun Wen, Zehao Zhu, Dejia Xu, Zhangyang Wang, et al. Lightgaussian: Unbounded 3d gaussian compression with 15x reduction and 200+ fps. *Advances in neural information processing systems*, 37: 140138–140158, 2025. 4

[2] Xianglong He, Junyi Chen, Sida Peng, Di Huang, Yangguang Li, Xiaoshui Huang, Chun Yuan, Wanli Ouyang, and Tong He. Gvgen: Text-to-3d generation with volumetric representation. In *European Conference on Computer Vision*, pages 463–479. Springer, 2024. 1

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems (NeurIPS)*, 33:6840–6851, 2020. 1

[4] Fangzhou Hong, Jiaxiang Tang, Ziang Cao, Min Shi, Tong Wu, Zhaoxi Chen, Shuai Yang, Tengfei Wang, Liang Pan, Dahua Lin, et al. 3dtopia: Large text-to-3d generation model with hybrid diffusion priors. *arXiv preprint arXiv:2403.02234*, 2024. 1

[5] 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. *arXiv preprint arXiv:2311.04400*, 2023. 1

[6] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 1, 3, 4

[7] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 1

[8] Yushi Lan, Fangzhou Hong, Shuai Yang, Shangchen Zhou, Xuyi Meng, Bo Dai, Xingang Pan, and Chen Change Loy. Ln3diff: Scalable latent neural fields diffusion for speedy 3d generation. In *European Conference on Computer Vision (ECCV)*, pages 112–130. Springer, 2024. 4

[9] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*, 2023. 1

[10] Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6526, 2024. 1

[11] Chenguo Lin, Panwang Pan, Bangbang Yang, Zeming Li, and Yadong Mu. Diffsplat: Repurposing image diffusion models for scalable 3d gaussian splat generation. In *International Conference on Learning Representations (ICLR)*, 2025. 1

[12] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9298–9309, 2023. 1

[13] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2837–2845, 2021. 1

[14] Qi Ma, Yue Li, Bin Ren, Nicu Sebe, Ender Konukoglu, Theo Gevers, Luc Van Gool, and Danda Pani Paudel. Shapesplat: A large-scale dataset of gaussian splats and their self-supervised pretraining. *arXiv preprint arXiv:2408.10906*, 2024. 2, 4

[15] 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 (CVPR)*, pages 4328–4338, 2023. 1

[16] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 3

[17] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4

[18] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 1

[19] 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 *International Conference on Machine Learning (ICML)*, pages 8748–8763. PmLR, 2021. 3

[20] Xuanchi Ren, Jiahui Huang, Xiaohui Zeng, Ken Museth, Sanja Fidler, and Francis Williams. Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4209–4219, 2024. 1

[21] Barbara Roessle, Norman Müller, Lorenzo Porzi, Samuel Rota Bulò, Peter Kontschieder, Angela Dai, and Matthias Nießner. L3dg: Latent 3d gaussian diffusion. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 1

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 1

[23] J Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20875–20886, 2023. 1

[24] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision (ECCV)*, pages 1–18. Springer, 2024. 1, 4

[25] 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*, pages 439–457. Springer, 2024. 1

[26] 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 *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1

[27] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 803–814, 2023. 4

[28] 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*, 2024. 1

[29] Haitao Yang, Yuan Dong, Hanwen Jiang, Dejia Xu, Georgios Pavlakos, and Qixing Huang. Atlas gaussians diffusion for 3d generation. *arXiv preprint arXiv:2408.13055*, 2024. 1

[30] Taoran Yi, Jiemin Fang, Junjie Wang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6796–6807, 2024. 1

[31] 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. 1

[32] Bowen Zhang, Yiji Cheng, Jiaolong Yang, Chunyu Wang, Feng Zhao, Yansong Tang, Dong Chen, and Baining Guo. Gaussiancube: Structuring gaussian splatting using optimal transport for 3d generative modeling. *arXiv e-prints*, pages arXiv–2403, 2024. 1

[33] Junsheng Zhou, Weiqi Zhang, and Yu-Shen Liu. Diffgs: Functional gaussian splatting diffusion. *Advances in Neural Information Processing Systems*, 37:37535–37560, 2025. 1