# A SURVEY ON TEXT TO 3D MODEL GENERATOR

**Arjun T Prakash[1], Arjun K S[2], Shijo Shaji[3], Sainath Raghunath[4], Rekha K S[56]**
[1-4]Student, [5]Assistant Professor
Department of CSE
College of Engineering, Kidangoor

arjuntprakash11@gmail.com[1], arjunkunnummel10@gmail.com[2], sshijo666@gmail.com[3], sainathrpn@gmail.com[4], rekhaks@ce-kgr.org[5]

***Abstract*—This survey investigates the use of Stable Diffusion models in the generation of 3D models, focusing on how text prompts can be transformed into 3D representations. The primary objective is to explore how a sequence of 2D images generated from a text description can be used to create a 3D point cloud, which serves as the basis for a fully realized 3D object. The survey also examines the fine-tuning process, implemented through Python-based techniques, that allows for the refinement of the model's shape, texture, and other attributes, providing users with the ability to customize the final output. A key aspect of this research is understanding how this approach simplifies 3D model creation compared to traditional methods, which typically require specialized software and extensive technical expertise. By enabling users to input text descriptions and automatically generate 3D models, the technique lowers the barrier to entry for individuals without advanced 3D modeling skills. The survey explores the process's advantages, such as the ease of customization through fine-tuning, which gives users greater control over the model's appearance. Moreover, the survey assesses the flexibility of this method in supporting various applications. The ability to export models in different formats allows integration into diverse fields, including game development, virtual reality, 3D printing, and architectural visualization. Ultimately, this survey seeks to evaluate how AI-driven 3D modeling can democratize the design process, empowering a broader audience to create complex and detailed 3D objects.**

## I. INTRODUCTION

The field of 3D modeling has long been dominated by specialized software and tools that require a high level of technical expertise to operate effectively. Traditional methods of 3D creation often involve complex workflows, including manual modeling, sculpting, and texturing, which can be both time-consuming and challenging for those without advanced skills. Moreover, the accessibility of these tools is often limited by the steep learning curve and the need for expensive hardware and software licenses. As a result, creating detailed and realistic 3D objects has remained a task reserved for professionals in fields like animation, gaming, architecture, and industrial design.

However, recent advancements in artificial intelligence, particularly in the area of generative models like Stable Diffusion, have the potential to transform this process. Stable Diffusion, an AI model known for its ability to generate highly detailed 2D images from text prompts, is now being explored for its ability to generate 3D representations. This opens up the possibility of generating complex 3D objects directly from textual descriptions, making the process more accessible to a wider audience. By converting text prompts into a sequence of 2D images, Stable Diffusion models can generate 3D point clouds, which serve as the starting point for creating 3D models. These point clouds can be refined and further customized through fine-tuning, enabling users to adjust the shape, texture, and other attributes of the model.

This approach eliminates many of the traditional barriers associated with 3D modeling, such as the need for advanced technical knowledge and specialized software. With AI-driven tools like Stable Diffusion, users can input simple text descriptions and, through the refinement process, produce high-quality 3D models tailored to their specific needs. This opens up new possibilities for creators across various industries, from gaming and virtual reality to 3D printing and architectural visualization. By enabling anyone with minimal technical expertise to generate and customize 3D objects, this method democratizes 3D modeling and empowers a broader range of individuals to create complex, realistic 3D assets. The potential for further advancements in AI-driven 3D modeling is vast, and this research seeks to explore how these innovations can reshape the way we think about and engage with 3D content creation.

## II. MOTIVATION

The rapid advancements in artificial intelligence and computer vision have opened up new possibilities in the field of 3D modeling. However, traditional 3D modeling techniques often require specialized software, extensive technical skills, and significant time investment. This can limit accessibility and creativity for many individuals and organizations. To address these limitations, this research aims to develop a novel approach to 3D model generation that is both accessible and efficient.

By leveraging the power of generative AI, specifically text-to-image models, we aim to democratize 3D model creation, empowering users to generate complex and realistic 3D models directly from textual descriptions. This innovative approach

has the potential to revolutionize various industries, including game development, film and animation, architecture and design, education and training, and e-commerce.

One of the primary motivations for this research is to reduce the barriers to entry for 3D modeling. By eliminating the need for specialized software and technical skills, we can empower individuals and organizations of all sizes to create high-quality 3D models. This can lead to increased creativity, innovation, and productivity across various industries.

Another key motivation is to accelerate the 3D model creation process. Traditional 3D modeling techniques can be time-consuming and labor-intensive, especially for complex models. By automating many of the steps involved in 3D model generation, we can significantly reduce the time and effort required to create high-quality models. This can lead to faster product development cycles, increased efficiency, and lower costs.

Furthermore, this research aims to improve the quality and realism of generated 3D models. By leveraging advanced machine learning techniques, we can generate highly detailed and accurate 3D models that are indistinguishable from real-world objects. This can lead to more immersive and realistic experiences in various applications, such as virtual reality and augmented reality.

, this research seeks to explore the potential of AI-powered 3D modeling to drive innovation and creativity. By enabling users to quickly and easily generate a wide range of 3D models, we can inspire new ideas, concepts, and designs. This can lead to the development of innovative products, services, and experiences that were previously unimaginable.

## III. LITERATURE SURVEY

Substantial progress has already been made in this field, providing a reference to understand the core principles and key concepts essential for this study.

**Poole, Ben, et al.** introduced Dreamfusion a novel approach for generating 3D models from text prompts by leveraging pretrained 2D diffusion models, bypassing the need for large-scale labeled 3D datasets typically required for 3D synthesis[1]. Traditional methods for text-to-3D generation are often constrained by limited 3D data availability and complex architectures specifically tailored to 3D model training. Instead, DreamFusion employs Score Distillation Sampling (SDS) to adapt a Neural Radiance Field (NeRF), which allows a 3D model to be optimized so that its 2D views match the characteristics of a given text prompt. This technique enables DreamFusion to produce high-quality 3D objects by repurposing 2D text-to-image diffusion models trained on extensive 2D image-text pairs, rather than relying on costly 3D model data. As a result, DreamFusion represents a significant advancement in making text-to-3D synthesis more accessible and efficient.

The ability of DreamFusion to create detailed and versatile 3D models solely from text descriptions opens up exciting possibilities for a range of creative industries. Since the model is based on 2D diffusion and NeRF rendering, it can synthesize 3D assets that can be viewed from multiple angles, illuminated in various ways, or integrated into different digital environments without extensive manual input or specialized training data. This approach has broad potential applications, from generating assets for video games and animations to streamlining workflows in visual design, VR, and AR. By enabling creators to produce realistic, dynamic 3D objects through simple text prompts, DreamFusion lowers the entry barrier for 3D content creation and provides an innovative tool for artists, developers, and other creative professionals in need of high-quality, customizable 3D assets.

**Rombach, Robin, et al.** introduced High-resolution image synthesis Proceedings of the IEEE/CVF conference on computer vision and pattern recognition[2]. 2022 introduces a novel approach to image synthesis by leveraging latent diffusion models (LDMs), which operate in a compressed, lower-dimensional latent space rather than the high-dimensional pixel space. Traditional diffusion models, while highly effective in generating realistic images, suffer from high computational costs due to their sequential denoising steps in pixel space, making them resource-intensive and less accessible. This paper addresses these limitations by proposing to train diffusion models within the latent space of pretrained autoencoders, which captures the essential visual information with reduced complexity. By separating the image representation into two stages—a perceptual compression phase that removes imperceptible details and a generative phase where the model focuses on semantic content—the authors achieve an optimal balance between computational efficiency and image quality. This approach not only reduces the number of parameters needed but also preserves high fidelity and detail, making it possible to generate high-resolution images while significantly cutting down on GPU requirements.

The architecture of the LDMs incorporates cross-attention layers to handle various conditioning inputs, such as text and spatial layouts, enhancing the flexibility of the model to perform multiple image generation tasks. The model demonstrates competitive performance across tasks like image inpainting, text-to-image synthesis, and super-resolution, setting new benchmarks in quality metrics while reducing both training and inference costs. By conditioning in the latent space and adding cross-attention for multimodal inputs, LDMs allow for high-resolution synthesis of images up to megapixel scale without compromising on quality or efficiency. The results show that LDMs are particularly well-suited for high-resolution tasks, delivering state-of- the-art results on multiple datasets, including CelebA-HQ, LSUN-Bedrooms, and ImageNet. The paper emphasizes the accessibility of LDMs due to their lower computational demands and releases the pretrained models, making them available for further research and application across various domains. This advance holds promise for broadening access to high-resolution image synthesis technology, particularly in settings with limited computational resources.

**Tian, Yonglong, et al.** introduced 3d shape programs presents an innovative method for reconstructing complex 3D shapes by introducing ”3D shape programs,” which integrate neural networks with symbolic representations to capture both detailed geometry and high-level structural features such as symmetry and repetition[3]. Traditional 3D modeling techniques typically rely on low-level representations like point clouds or voxel grids, which can struggle with capturing structural regularities. In contrast, this method combines bottom-up recognition with top-down symbolic programming, where shapes are described through primitives and operations that enforce structure. The model infers 3D shape programs from raw, unannotated 3D data and can execute these programs to reconstruct the shapes. This approach offers a new domain-specific language (DSL) for representing shapes, allowing the system to build a structured understanding of shape attributes like size, orientation, and repetition in the object’s parts.

A significant feature of the method is its self-supervised learning framework, which avoids the need for extensive labeled data by learning through reconstruction errors. The model uses a differentiable neural program executor, which enables fine-tuning through gradient back-propagation to continuously improve shape fidelity and structural coherence. Experimental results show that this approach achieves high accuracy in reconstructing diverse, intricate shapes from categories such as furniture and objects, while also adapting well to unseen classes. The model’s ability to infer 3D shapes from images, reconstruct stable and connected 3D forms, and operate without shape annotations highlights its potential for applications in 3D modeling, graphics, and computer vision.

**Tang, Junshu, et al.** introduced Make-it-3d Creating high-quality 3D models from a single 2D image has been a longstanding challenge in computer vision, as it requires accurately reconstructing both the geometry and textures of an object[4]. Traditional methods often rely on multiple images or are restricted to specific object categories, while recent advancements with techniques like neural radiance fields (NeRFs) still struggle with generalizing across different object types and often require multi-view data. This limits their practical applications. ”Make-It-3D” overcomes these challenges by leveraging diffusion models, which can encode 3D knowledge from 2D images. The method uses a two-stage pipeline: the first stage generates a rough 3D model using NeRF optimized with score distillation sampling, guided by the input image and multi-view supervision from the diffusion model. While this first stage captures the object’s geometry, it lacks texture fidelity. In the second stage, Make-It-3D refines the model by creating a textured point cloud, transferring high-quality textures from the reference image and enhancing unseen areas through diffusion-based rendering. This approach achieves both geometric accuracy and realistic textures, offering flexibility for tasks like texture editing and text-to-3D generation. Make-It-3D sets a new standard for high-fidelity 3D modeling from a single image, with applications in fields like virtual reality, gaming, and digital content creation.

**Zhan, Guanqi,et al.** introduced general protocol The Thirty-eighth Annual Conference on Neural Information Processing Systems[5]. 2023 presents a framework to evaluate how well large-scale vision models, such as CLIP, DINO, and Stable Diffusion, understand 3D scene properties from 2D images. The authors propose a three-step probing protocol: selecting datasets with ground truth annotations, identifying optimal feature representations, and training classifiers to predict specific properties.

The study reveals that models like Stable Diffusion and DINOv2 demonstrate a strong understanding of scene geometry, support relations, shadows, and depth. However, they struggle with occlusion and material recognition. These findings have significant implications for various applications, including synthetic image detection, augmented reality, and model refinement.

**Beltramelli, Tony, et al.** introduced pix2code Proceedings of the ACM SIGCHI symposium on engineering interactive computing systems[6]. 2018 introduces pix2code, a deep learning model that automates the process of generating code from GUI screenshots. The model combines CNNs and LSTMs to process GUI images and generate code in a custom DSL, which can be compiled into target platform languages. The researchers synthesized datasets for training, and the model achieved over 77 percentage accuracy across iOS, Android, and web platforms. This research demonstrates the potential of automating GUI code generation, especially for web-based applications. By leveraging the power of deep learning, pix2code has the potential to significantly reduce development time and effort, making it a valuable tool for developers and designers. This advancement could revolutionize the way user interfaces are created, allowing designers to focus on the visual aspects while the model generates the underlying code, leading to faster and more efficient development cycles.

**Xu, Jiale, et al.** introduced Dream3d Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition[7]. 2023 introduces Dream3D, a framework for generating high-quality 3D models from text prompts. Dream3D addresses limitations in prior methods by incorporating 3D shape priors, which provide a structured base for optimization. The pipeline consists of two stages: text-to-shape generation and CLIP guidance. The first stage produces an initial 3D shape prior, and the second stage refines it to match the text prompt. Dream3D demonstrates significant potential for text-driven 3D content creation, improving upon previous methods in terms of accuracy, visual quality, and efficiency. By leveraging the power of text-to-image diffusion models and 3D shape priors, Dream3D paves the way for innovative applications in fields

such as game development, virtual reality, product design, and architectural visualization. This framework has the potential to democratize 3D content creation, enabling users to generate complex 3D models from simple text descriptions.

**Liu, Zhengzhe** introduced text-guided 3d shape generation proposes a framework for creating high-fidelity 3D shapes from text descriptions[8]. By introducing a decoupled shape and color prediction, a word-level spatial transformer, and a style-based IMLE generator, the framework improves text-shape alignment, diversity, and overall quality. The framework demonstrates superior performance on existing benchmarks and enables text-based shape manipulation, allowing users to modify the generated 3D shapes by editing the text input. This research advances the state-of-the-art in text-to-3D generation, opening up new possibilities for efficient and creative 3D content creation across various applications, including game development, virtual reality, product design, and architectural visualization.

**Tewari, Ayush, et al.** introduced State of the art on neural rendering Neural rendering merges traditional rendering techniques with deep learning to generate realistic images, animations, and 3D models[9]. Unlike traditional methods, neural rendering bypasses manual scene construction and directly produces realistic outputs using generative models. This approach enables the simulation of complex physical properties like lighting changes, textures, and viewpoints, even handling dynamic elements with fewer data constraints. This field has revolutionized image and video synthesis, offering interactive and controlled scene creation for applications in entertainment, virtual and augmented reality, telepresence, and digital human avatars. Neural rendering bridges the gap between physical realism and computational efficiency, making high-quality graphics more accessible and versatile.

**Xu, Qun-Ce, et al.** introduced deep learning-based 3D shape generation delves into the intersection of 3D shape generation and deep learning, highlighting the challenges posed by 3D representations compared to 2D images[10]. The survey categorizes deep learning-based 3D shape generation methods by representation and generative architecture, exploring techniques like GAN-based models, VAEs, flow-based models, and autoregressive models. Each approach is evaluated for its ability to capture geometric details, handle various topologies, and efficiently manage 3D data complexity. The paper concludes by discussing open challenges, such as the need for more adaptable architectures and larger datasets, and suggests future research directions, including the development of hybrid models that combine the strengths of multiple techniques, the exploration of new 3D representations that are more suitable for deep learning, and the integration of semantic information to guide the generation process.

**Lee, Hanhung, Manolis Savva, et al.** introduced Text-to-3D Shape Generation explores the state-of-the-art in text-to-3D shape generation, focusing on methods that leverage text prompts to create 3D content[11]. The report categorizes approaches into four main families: 3DPT (paired 3D and text data), NO3D (no 3D data), hybrid approaches, and foundational techniques like NeRF and Diffusion Models.3DPT methods achieve high-quality output by leveraging supervised learning, though they are limited by the dataset's scope. NO3D methods often employ prompt-specific optimizations using differentiable renderers to maximize similarity between text and generated images, although such techniques can be computationally intensive. Hybrid approaches combine text-to-image and image-to-3D steps to enhance consistency across views and scale across diverse scenes.The report also addresses significant challenges in achieving high-quality 3D generation, evaluating text-to-3D quality, and identifying future directions, including refining unsupervised learning strategies and enabling intuitive shape editing for users.

## IV. CONCLUSION

In conclusion, this survey has explored the state-of-the-art advancements in text-to-3D shape generation and GUI code generation. We have discussed various techniques, challenges, and future directions in these fields.

Text-to-3D shape generation has made significant strides, with models like Dream3D demonstrating impressive results in generating high-quality 3D shapes from text descriptions. However, challenges remain in terms of generating complex shapes, handling diverse text prompts, and ensuring consistency between the text and the generated 3D model.

GUI code generation, as exemplified by pix2code, has shown promising results in automating the process of converting GUI designs into code. However, further research is needed to improve accuracy, handle complex layouts, and support a wider range of platforms and design styles.

As AI continues to advance, we can expect further breakthroughs in these areas. Future research directions may include exploring more sophisticated generative models, incorporating semantic understanding, and developing robust evaluation metrics. By addressing these challenges and pushing the boundaries of AI, we can unlock the full potential of text-to-3D shape generation and GUI code generation, leading to more efficient and creative design processes.

## REFERENCES


[1] Poole, Ben, et al. ”Dreamfusion: Text-to-3d using 2d diffusion.” arXiv preprint arXiv:2209.14988 (2022).

[2] Rombach, Robin, et al. ”High-resolution image synthesis with latent diffusion models.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022

[3]Tian, Yonglong, et al. ”Learning to infer and execute 3d shape programs.” arXiv preprint arXiv:1901.02875 (2019).

[4]Tang, Junshu, et al. ”Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior.” Proceedings of the IEEE/CVF international conference on computer vision. 2023.

[5]Zhan, Guanqi, et al. ”A general protocol to probe large vision models for 3d physical understanding.” The Thirty-eighth Annual Conference on Neural Information Processing Systems. 2023.

[6] Beltramelli, Tony. ”pix2code: Generating code from a graphical user interface screenshot.” Proceedings of the ACM SIGCHI symposium on engineering interactive computing systems. 2018.

[7] Xu, Jiale, et al. ”Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[8] Liu, Zhengzhe, et al. ”Towards implicit text-guided 3d shape generation.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[9] Tewari, Ayush, et al. ”State of the art on neural rendering.” Computer Graphics Forum. Vol. 39. No. 2. 2020.

[10] Xu, Qun-Ce, Tai-Jiang Mu, and Yong-Liang Yang. ”A survey of deep learning-based 3D shape generation.” Computational Visual Media 9.3 (2023): 407-442.

[11] Lee, Hanhung, Manolis Savva, and Angel X. Chang. ”Text-to-3D Shape Generation.” Computer Graphics Forum. 2024.