Speech-2-Sculpt: Transforming Ideas into 3D Printed Creations

Trushant Adeshara
Department of Robotics
University of Michigan
Ann Arbor, USA
trushant@umich.edu

Pannaga Sudarshan
Department of Robotics
University of Michigan
Ann Arbor, USA
pannaga@umich.edu

Kajal Awasthi
Department of Robotics
University of Michigan
Ann Arbor, USA
kajalaw@umich.edu

Saket Pradhan
Department of Robotics
University of Michigan
Ann Arbor, USA
saketp@umich.edu

***Abstract*—We introduce ”Speech-2-Sculpt”; an approach is designed for efficient 3D mesh generation from speech input utilizing advanced reconstruction models. Our implementation adapts the InstantMesh framework to generate printable meshes using speech as input, and compares its performance with other methods of 3D printing objects – using traditional CAD software, or existing generative models such as *Point-E* or *Point2Mesh*. The experimental results highlight the overall efficiency in terms of quality and ease of printing an object from the ground up via different techniques. This work contributes to the broader field of 3D generative AI by incorporating a speech-driven, CAD-software-free approach to 3D printing. Our findings underscore the potential of advanced reconstruction models in practical applications, offering substantial improvements over traditional methods both in speed and reasonable output quality. We use OpenAI's Whisper to convert the spoken input into a textual prompt that in-turn conditions our stable diffusion model to generate multi-view images representing the desired 3D object. These images then serve as input to the InstantMesh framework, which employs a sparse-view large reconstruction model based on the LRM architecture to predict a high-fidelity 3D mesh optimized for 3D printing. By combining cutting-edge generative models, our pipeline bypasses traditional CAD software, making sophisticated 3D modeling accessible to non-technical users.**

***Index Terms*—3D printing, CAD, diffusion, NeRF, Point-E, Point2Mesh, 3D reconstruction**

I. Introduction

A cornerstone of modern additive manufacturing– 3D printing, widely known as 3D printing, has emerged as a transformative technology that fundamentally alters how products are designed, developed, and produced. This process constructs objects layer by layer, allowing for the creation of complex structures that were previously unattainable with traditional manufacturing techniques. The advent of 3D printing has not only optimized material usage and streamlined production processes but has also introduced unprecedented flexibility in customization and rapid prototyping [1].

The convergence of natural language processing (NLP) and 3D modeling technologies has revolutionized digital fabrication, enabling the direct transformation of textual descriptions into tangible objects. These technological advancements have unlocked unprecedented opportunities, allowing for the seamless integration of NLP, vision transformers, and other sophisticated tools to develop innovative pipelines in ways previously unimaginable. This democratization of 3D printing technology empowers both enthusiasts and professionals by making sophisticated design and production techniques more accessible.

Further enhancing this progression, the incorporation of machine learning algorithms and computational design has facilitated the automation of complex modeling tasks that previously required extensive manual input and specialized skills. As a result, there is now a broader scope for creativity and experimentation in design across various sectors, including healthcare, architecture, and consumer goods. In addition to drastically aiding in streamlining production processes, it also promotes a more inclusive environment where anyone can turn their ideas into physical reality without needing technical expertise in traditional 3D modeling software.

In this project, we aim to take this democratization a step further by introducing Speech-2-Sculpt, a pioneering pipeline that enables the direct transformation of spoken language into 3D printable objects. By leveraging state-of-the-art speech recognition, text-to-image generation, and image-to-3D reconstruction models, our approach effectively bypasses the need for traditional 3D design software, making the creation of 3D assets accessible to an even broader audience.

At the core of our approach lies the InstantMesh framework [2], a cutting-edge feed-forward model designed for efficient 3D mesh generation directly from a single image. Our key innovation involves adapting this framework to accept multi-view images generated by a stable diffusion model conditioned on text prompts derived from the user's spoken input. By integrating OpenAI's robust Whisper speech recognition model [3], we accurately capture the nuances of spoken language, ensuring that the resulting 3D models faithfully reflect the user's intent and specifications.

The research problem addressed here is formulated as follows: **Given a speech input**, we **leverage the InstantMesh model** as our backbone to generate a multi-view representation of the desired object **based on the spoken description**, and subsequently **generate a mesh object file** suitable for 3D printing. **We assume** that the model provides a reasonably accurate 3D reconstruction of the object from the multiple unique views generated by the multi-view diffusion model.

II. Related Work

Converting Images to 3D models: Early endeavors in translating images to 3D primarily concentrated on single-view reconstruction tasks [2]. With advancements in diffusion models, promising studies have explored 3D generative modeling conditioned on images across various formats such as point clouds, meshes, SDF grids [4], and neural fields [5]. Although these techniques show considerable promise, their application to open-world objects remains challenging due to the constraints imposed by the limited amount of training data available.

Stable Diffusion: For our text-to-image conversion pipeline, the underlying model uses a stable diffusion model [6]. Stable Diffusion is a text-to-image diffusion model developed by Stability AI. It is capable of generating high-quality images from textual descriptions. The model is based on the diffusion probabilistic model, which iteratively refines an initial random noise tensor to produce the final image. Stable Diffusion has been trained on a large dataset of image-text pairs, allowing it to learn the complex relationships between visual and textual representations. This model has gained widespread popularity due to its ability to generate diverse and creative images with high fidelity.

Point-E: Point-E [4] is a system designed to generate 3D point clouds from textual prompts. It uses a transformer encoder to process the text prompt and a transformer decoder to auto regressively generate the 3D point coordinates. During training, it is optimized to maximize the likelihood of the point cloud given the text prompt under a Gaussian kernel density estimation objective. The key innovation in Point-E is the use of vector quantized Gaussian attention, which allows it to model the complex distributions of 3D point clouds and capture fine-grained geometric details described in the text prompts. Point-E takes a text prompt as input and produces a set of 3D coordinates representing the desired object. This point cloud can be further processed or converted into a mesh representation for various applications, such as 3D printing or visualization.

Point2Mesh: Point2Mesh [7] is a self-prior technique for deformable mesh reconstruction from point clouds. It aims to generate a high-quality mesh representation that accurately captures the underlying geometry of a given point cloud. starts with an initial template mesh and iteratively deforms it to fit the input point cloud using a Graph Convolutional Network (GCN) [8]. The GCN operates directly on the mesh edges, enabling efficient convolutional operations while respecting the mesh topology. During training, Point2Mesh minimizes a combined loss function that considers both the point cloud fitting error and regularization terms that encourage smooth deformations and preservation of the initial mesh topology, resulting in high-quality meshes that conform to the input point clouds.The method employs a self-prior regularization term that encourages the output mesh to maintain a natural and plausible shape.

NeRF-guided 3D Reconstruction: Neural Radiance Fields [5] is a technique for representing and rendering 3D scenes from a set of 2D images. NeRF represents the radiance field as a fully-connected neural network that takes 3D coordinates and 2D viewing directions as inputs and outputs the volume density and view-dependent radiance at that spatial location. It is trained on images with known camera poses by optimizing a volumetric rendering loss that compares the rendered images to the ground truth. They use positional encoding to map the input 3D coordinates into a higher-dimensional space, which allows the neural network to better capture high-frequency details and complex geometry in the scene representation. This has enabled impressive view synthesis and 3D reconstruction results from a relatively compact network architecture.

III. Algorithmic Extension

Given the recent advancements in generative AI models for 3D reconstruction as mentioned in section II, we aim to explore an innovative approach that seamlessly integrates speech recognition, text-to-image generation, and image-to-3D conversion. Our objective was to develop an end-to-end pipeline that empowers users to create 3D printable objects directly from spoken descriptions, eliminating the need for traditional 3D modeling software or technical expertise.

Initially, we attempted to implement InstructP2P [2], a method that generates 3D point clouds from text prompts. However, due to the unpublished nature of the work, reproducing the results proved challenging. Additionally, as highlighted in InstantMesh, radiance field-based mesh construction techniques tend to achieve superior results compared to point cloud-based methods. Consequently, we pivoted our approach to leverage the promising diffusion model-based framework introduced in InstantMesh.

Our proposed algorithmic extension involves integrating a speech-to-text model, specifically OpenAI's Whisper [3], at the front-end of the InstantMesh pipeline. The spoken input is first converted into a textual prompt using Whisper's robust speech recognition capabilities. This text prompt is then passed to a stable diffusion model [2], [5], [9]–[12], which generates a set of multi-view images representing the desired 3D object. These generated images serve as input to the InstantMesh framework, which utilizes a sparse-view large reconstruction model to produce a high-quality 3D mesh suitable for 3D printing. By combining these state-of-the-art components, our pipeline enables users to transform their spoken ideas into tangible 3D printed creations in an efficient and accessible manner.

IV. Methodology

In this study, we explore the transformation of speech into 3D printed objects through two innovative pipelines as shown in Figure 1, both commencing with the conversion of a spoken input into a text prompt utilizing OpenAI's Whisper model [3], a robust speech-to-text transformation tool. This model is known for its accuracy in speech-to-text conversion, providing a reliable textual representation of verbal descriptions. The methodologies diverge after this initial step, incorporating different technologies to to yield 3D printable meshes.

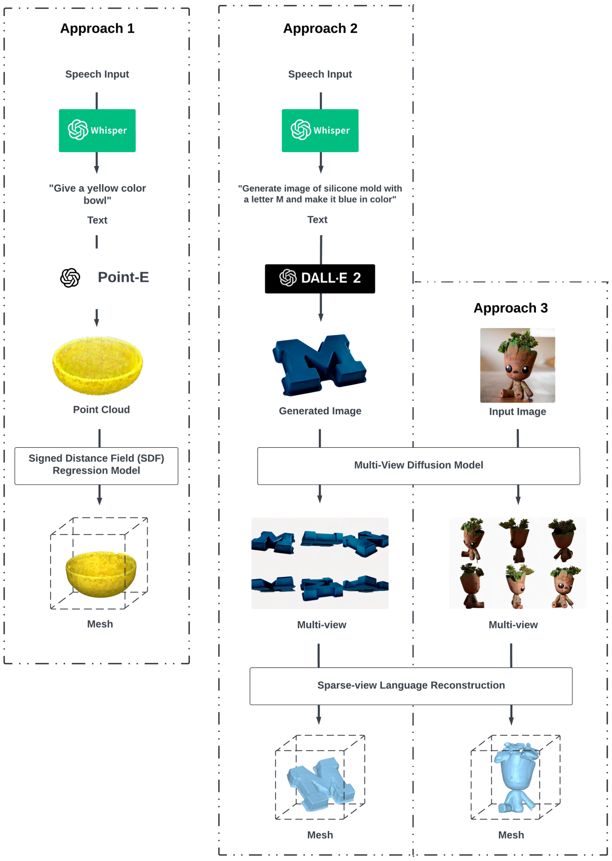
Approach 1
Speech Input
Whisper
"Give a yellow color bowl"
Text
Point-E
Point Cloud
Signed Distance Field (SDF) Regression Model
Mesh
Approach 2
Speech Input
Whisper
"Generate image of silicone mold with a letter M and make it blue in color"
Text
DALL·E 2
Generated Image
Approach 3
Input Image
Multi-View Diffusion Model
Multi-view
Multi-view
Sparse-view Language Reconstruction
Mesh
Mesh

Fig. 1. Speech-2-Sculpt end-to-end flow

The first approach bypasses image generation and directly transforms the text prompt into a point cloud where it employs the Point-E model [4]. The Point-E model is specifically designed for this purpose, utilizing a neural network that predicts a set of 3D coordinates (points) based on textual input.

These coordinates collectively form a point cloud that approximates the shape and size of the desired object. The point cloud is then converted into a mesh using a specialized Point-to-Mesh [7] conversion process. This approach learns from a single object, by optimizing the weights of a CNN to deform some initial mesh to shrink-wrap the input point cloud. This step ensures that the discrete points are seamlessly connected to form a continuous, printable 3D structure that maintains the geometric accuracy and integrity of the design described in the text prompt.

Conversely, the first approach utilizes a Stable Diffusion-based model [9]–[13], an advanced deep learning model renowned to convert the text prompt into a detailed image that visually interprets the input with subsequent mesh generation. Stable Diffusion excels in creating coherent and contextually appropriate images from textual data.

Following image synthesis, the pipeline employs an Instant Mesh model [2] to generate multiple views of the synthesized image from diverse angles. This multi-view approach is essential for developing a comprehensive 3D mesh that captures various facets of the image.The multi-view images are then used to construct a final mesh model of the object. This mesh is optimized for 3D printing, providing the necessary detail and structure integrity required for physical realization.

Both methodologies aim to efficiently bridge the gap between verbal commands and physical objects through advanced AI-driven processes. By leveraging different techniques—image synthesis followed by mesh generation in the first method, and direct point cloud creation followed by mesh conversion in the second—this research explores varied pathways to achieve high-quality 3D printed outcomes from spoken language inputs. This integrative approach not only highlights the capabilities and flexibilities of modern 3D printing technology but also pushes the boundaries of how artificial intelligence can interact with and enhance creative and manufacturing processes.

V. Experiments and Results

A. Experimental Setup

In our project, we compare two different pipelines. From Figure 1, we note that both the models take in speech as input and utilize whisper model to obtain the text description of the same. The whisper model generates accurate descriptions of a given speech input, which is utilized by point-E network (*Approach 1*) to generate a point cloud, $\in R^3$, which is further converted to a mesh file by leveraging SDF. Further, *Approach 2* utilizes the text description from the whisper model to obtain text description, and we feed this into a diffusion model [14], that generates an image according to the text description. This image is now treated as the input for the InstantMesh model. Although this approach outputs a mesh file similar to *Approach 1*, the underlying architecture or backbone of InstantMesh is unlike any other exisiting methodology, wherein the mesh is actually generated after reconstructing the object in the scene using NeRF.

B. Results

The outcomes of our experimental pipelines are detailed below:

Speech to Text Conversion: The ASR model, specifically OpenAI's Whisper [3], demonstrated high accuracy in converting speech to text, achieving between 95% and 98.5% accuracy. This performance underpins the reliability of the initial step in our processing chain.

Point Cloud Generation (Approach 1): As depicted in 1, the first pipeline utilizes the Point-E model to convert text prompts into point clouds. This model successfully generates sufficiently detailed point clouds within a timeframe of only 1-2 minutes, indicating both efficiency and effectiveness.

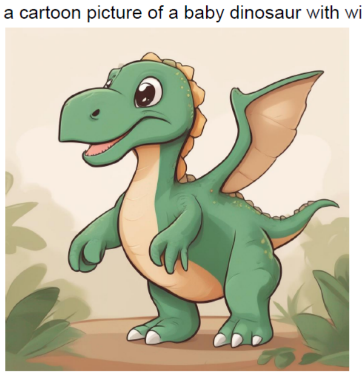

Fig. 2. Multiview images of Cartoon Dinosaur

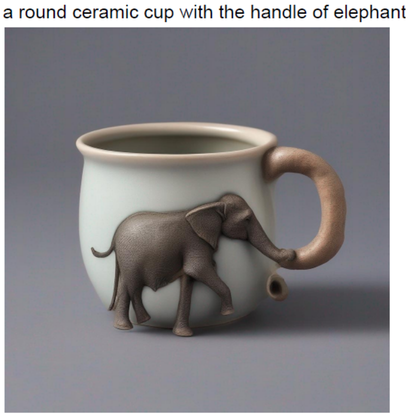

Fig. 3. Multiview images of Cup with elephant handle

Image and Mesh Generation (Approach 2): Employing the DALL-E2 model [14] by OpenAI, this approach transforms text prompts into detailed images as shown in Figure 2, 3. Subsequently, it generates multiview representations of these images as shown in Figure 4, 5 to facilitate mesh file creation. A parallel pathway within this approach also processes direct image inputs to produce comparable mesh structures, using multiview images as an intermediate step. This method showcases versatility in handling both text and image inputs effectively.

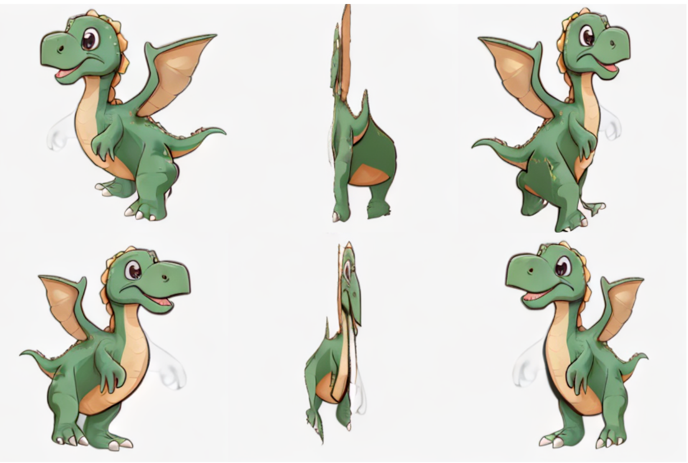

Fig. 4. Cartoon Dinosaur with speech input

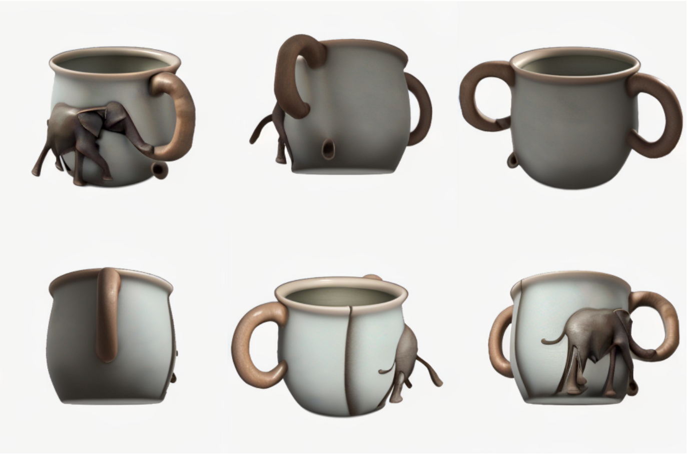

Fig. 5. Cup with elephant handle with speech input

Mesh Quality Comparison: The mesh structures generated using the SDF model in the first approach did not exhibit the same level of quality as those produced in Approach 2. This observation is based on the physical characteristics of the printed meshes, where those from Approach 2 displayed superior detail and fidelity.

Following *Approach 2* we generated a mesh file of objects as shown in Figures 3, and 2. Figures 6, and 7 depict the 3D printed objects from the mesh files we obtain.

VI. Discussions

A. Limitations

We encountered several challenges with the models used in our pipeline, particularly with the InstantMesh and Point-E models. InstantMesh demonstrated difficulties in handling complex textures and was notably limited to objects that are watertight or concave. For instance, when attempting to model a cup, InstantMesh could not accurately represent the depth as shown in Figure 6, resulting instead in a superficial depiction of depth. This model also struggled with complex prompts, occasionally producing images that lacked relevance

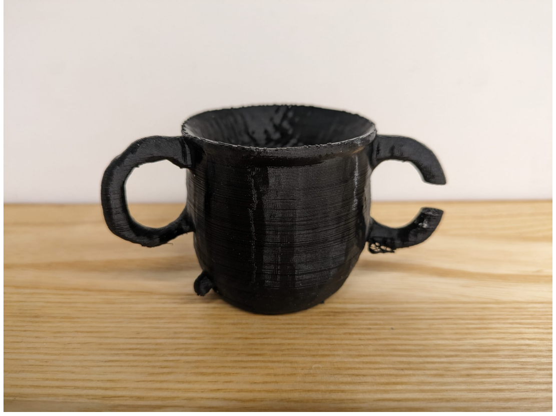

Fig. 6. Cup with elephant handle

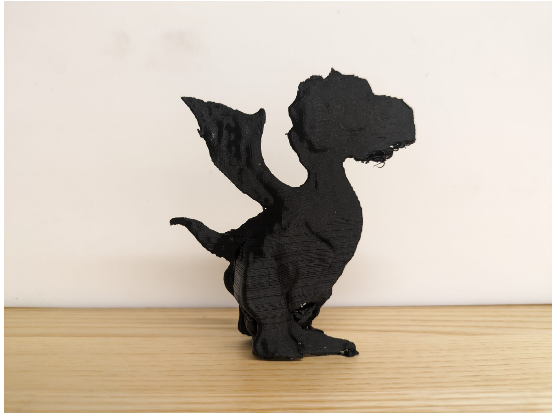

Fig. 7. Cartoon Dinosaur

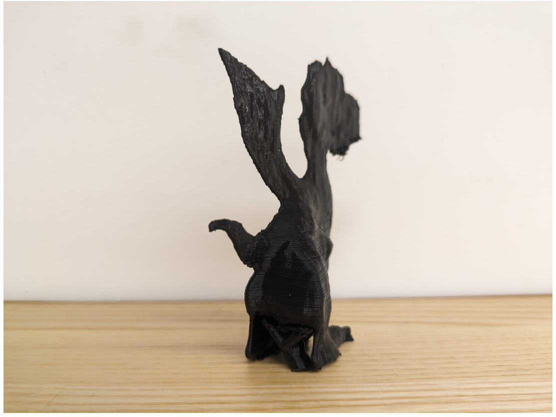

Fig. 8. Incorrect Depth Estimate from the 2D image

to the intended outputs.

Additionally, our approach lacks a quantitative metric for measuring changes in geometry in some cases, example in the cartoon dinosaur as shown in Figure 3, which is crucial for precise model evaluation. The diffusion models used, despite their versatility, failed to capture depth information adequately when processing images from sketches (a 2D picture), often resulting in the generation of flat meshes.

Point-E's normalization of point density further compounded these issues, as it reduced the granularity necessary for producing high-quality meshes, thereby affecting the overall fidelity and utility of the generated 3D models. These limitations highlight significant areas for future research and development, particularly in improving texture synthesis, depth perception, and geometric accuracy in our modeling processes.

B. Future Work

While our Speech-2-Sculpt pipeline demonstrates promising results in generating 3D printable meshes from spoken input, there remain several opportunities for further enhancement and exploration. One area for improvement lies in increasing the resolution of the generated 3D meshes. Currently, our pipeline inherits the resolution limitations of the InstantMesh framework, which produces 64x64 triplanes as the 3D representation. Exploring architectures capable of generating higher-resolution 3D representations could unlock the potential for more intricate and high-definition 3D modeling. Additionally, the multi-view inconsistency inherent in the diffusion model used for image generation can introduce artifacts or inconsistencies in the final 3D mesh. Investigating advanced multi-view diffusion architectures or incorporating techniques to enforce greater consistency could lead to improved fidelity and coherence in the generated 3D assets.

Furthermore, while our current implementation leverages the FlexiCubes module for efficient mesh extraction, this approach may struggle to accurately capture intricate thin structures or fine details. Exploring alternative differentiable surface extraction methods or hybrid approaches combining the strengths of neural radiance fields and explicit mesh representations could enhance the preservation of intricate geometric details in the final output.

VII. Conclusion

In this work, we present Speech-2-Sculpt, a novel end-to-end pipeline that enables the transformation of spoken ideas into 3D printed objects. By seamlessly integrating speech recognition, text-to-image generation, and image-to-3D reconstruction technologies, our framework empowers users to create tangible 3D assets directly from verbal descriptions. It allows for the creation of 3D assets from single-view images, which opens up a myriad of possibilities across diverse fields such as virtual reality, industrial design, gaming, and animation.

At the core of our approach lies the InstantMesh framework, which we have adapted to accept multi-view images generated by a stable diffusion model conditioned on text prompts. These text prompts are derived from the spoken input using OpenAI's Whisper speech recognition model. This innovative integration allows us to bypass the need for traditional 3D modeling software, making the creation of 3D printable objects more accessible to a broader audience.

Our experimental results demonstrate the effectiveness of our pipeline in generating high-quality 3D meshes suitable for 3D printing. The generated meshes exhibit plausible geometries and textures, capturing the intricate details described in the spoken input.

Our pipeline not only streamlines the 3D printing process but also promotes inclusivity by democratizing access to sophisticated design and fabrication techniques. With our framework, individuals without extensive technical expertise in 3D modeling can effortlessly transform their creative visions into tangible objects, fostering innovation and creativity across various domains. Future research could focus on enhancing the resolution and fidelity of the generated meshes, addressing multi-view inconsistencies, and potentially incorporating interactive feedback mechanisms to refine the outputs iteratively.

References

[1] M. Yampolskiy, T. R. Andel, J. T. McDonald, W. B. Glisson, and A. Yasinsac, "Towards security of additive layer manufacturing," 2015.

[2] J. Xu, W. Cheng, Y. Gao, X. Wang, S. Gao, and Y. Shan, "Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models," 2024.

[3] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[4] A. Nichol, H. Jun, P. Dhariwal, P. Mishkin, and M. Chen, "Point-e: A system for generating 3d point clouds from complex prompts," 2022.

[5] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," 2020.

[6] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," 2023.

[7] R. Hanocka, G. Metzer, R. Giryes, and D. Cohen-Or, "Point2mesh: a self-prior for deformable meshes," *ACM Transactions on Graphics*, vol. 39, no. 4, Aug. 2020. [Online]. Available: http://dx.doi.org/10.1145/3386569.3392415

[8] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2017.

[9] J. Ho and T. Salimans, "Classifier-free diffusion guidance," 2022.

[10] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2022.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021.

[12] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," 2022.

[13] A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres, "Quantifying the carbon emissions of machine learning," 2019.

[14] Z. Shi, X. Zhou, X. Qiu, and X. Zhu, "Improving image captioning with better use of captions," *CoRR*, vol. abs/2006.11807, 2020. [Online]. Available: https://arxiv.org/abs/2006.11807