

# 3D-PREMISE: CAN LARGE LANGUAGE MODELS GENERATE 3D SHAPES WITH SHARP FEATURES AND PARAMETRIC CONTROL?

**Zeqing Yuan\***

Zhejiang University

**Haoxuan Lan\***

Zhejiang University

**Qiang Zou**

Zhejiang University

**Junbo Zhao**

Zhejiang University

## ABSTRACT

Recent advancements in implicit 3D representations and generative models have markedly propelled the field of 3D object generation forward. However, it remains a significant challenge to accurately model geometries with defined sharp features under parametric controls, which is crucial in fields like industrial design and manufacturing. To bridge this gap, we introduce a framework that employs Large Language Models (LLMs) to generate text-driven 3D shapes, manipulating 3D software via program synthesis. We present 3D-PreMise, a dataset specifically tailored for **3D Parametric Modeling of industrial shapes**, designed to explore state-of-the-art LLMs within our proposed pipeline. Our work reveals effective generation strategies and delves into the self-correction capabilities of LLMs using a visual interface. Our work highlights both the potential and limitations of LLMs in 3D parametric modeling for industrial applications.

## 1 INTRODUCTION

The integration of software in computational design and manufacturing has revolutionized various engineering sectors, offering significant advancements and efficiencies. However, this integration necessitates a deep understanding of domain-specific knowledge, underscoring the necessity for the development of more accessible interfaces, such as natural language processing, in industrial applications.

In recent years, generative models have made remarkable strides in non-engineering domains. The field of 3D object generation is undergoing a transformative shift, propelled by groundbreaking developments in implicit 3D representation (Mildenhall et al., 2020; Kerbl et al., 2023) and advanced generative models (Ho et al., 2020). While current methods employing implicit representations can generate intricate objects from text and image inputs, they inherently cannot preserve sharp features that contains engineering semantics (e.g. sharp edges, precise dihedral angles) when converting to explicit representations, a critical limitation for applications such as industrial design and manufacturing. For instance, generating a chair with precise dimensions or a plastic bottle cap with parallel and equidistant edges remains a challenge.

This work explores a novel approach to 3D object generation merging the capabilities of LLMs with the precision of 3D software. We harness LLMs to interpret text inputs and generate code to control 3D modeling software like Blender, producing compact 3D shapes with sharp features that adhere to strict parametric controls. This pipeline is advantageous for two reasons:

Firstly, 3D modeling code emerges as an ideal medium for depicting everyday objects, particularly in industrial settings, due to its compactness and high controllability that align well with design intentions. In contrast, implicit 3D representations such as neural radiance fields and Gaussian Splatting, despite their expressiveness for intricate shapes, struggle with maintaining sharp features due to inevitable conversion. Furthermore, directly generating explicit representations, like meshes, currently cannot offer precise parametric controls.

Secondly, advanced LLMs have shown remarkable understanding of 3D spaces and aptitude in programming. They can perform 3D spatial reasoning, planning (Sun et al., 2023) and demonstrate

\*equal contribution. Zeqing is the project lead. Haoxuan participated in dataset construction and testing.

---

high accuracy in programming tasks (Zhou et al., 2023). Therefore, it is promising that, given appropriate prompts and fine-tuning, LLMs can effectively generate 3D objects using modeling code as an intermediary.

Despite its potential, this approach encounters significant challenges. Firstly, the model must adeptly reason about spatial relationships and cohesively assemble object components within a unified coordinate system. Additionally, it entails intricate geometric calculations, including the application of complex trigonometric functions. Lastly, the LLM must possess commonsense reasoning to deduce stable and realistic structures from brief natural language descriptions.

To validate this approach, we introduce 3D-PreMise, a dataset comprising test programs and data samples with problem descriptions and corresponding ground-truth code. This dataset focuses on typical industrial objects with exact geometries. We conduct experiments with state-of-the-art LLMs to evaluate their performance and analyze generation strategies.

Our contributions are three-fold:

- We introduce a self-refining framework for 3D shape generation with parametric control, which leverages LLMs to control 3D software through code;
- We construct a benchmark dataset and conduct experiments to analyze the capacities of cutting-edge LLMs;
- We explore effective generation strategies and self-correcting capacity through a multi-modal interface.

## 2 RELATED WORK

### 2.1 TEXT-DRIVEN 3D OBJECT GENERATION

The field of Text-driven 3D object generation has witnessed remarkable progress. Clip-Mesh (Khalid et al., 2022) utilizes CLIP guidance for zero-shot 3D generation. DreamFusion (Poole et al., 2022) proposes a stable paradigm using Score Distillation Sampling (SDS) loss to distill pre-trained text-to-image diffusion models and achieves significant improvement. Magic3D (Lin et al., 2022) enhances visual quality in high resolution by introducing a two-stage optimization framework. Fantasia3D (Chen et al., 2023) introduces richer representations. ProlificDreamer (Wang et al., 2023) proposes variational score distillation (VSD) to address the over-saturation problem. DreamBooth3D (Raj et al., 2023) proposes a 3-stage optimization strategy to jointly leverage 3D consistency of NeRF and the personalization capability of text-to-image models. MVDream (Shi et al., 2023) improves the geometric consistency by applying multi-view diffusion model as a 3D prior with Score Distillation. To accelerate generation process, Instant3D (Li et al., 2023) proposes a two-stage paradigm that first generates a sparse set of views and then regresses NeRF with a transformer-based reconstructor; DreamGaussian (Tang et al., 2023) designs a generative 3D Gaussian Splatting model for faster generation speed.

However, these methods rely on implicit 3D representations, which inherently struggles to retain sharp features embodying engineering semantics during the conversion to practical explicit forms. Consequently, despite their increasing sophistication, these methods fall short in meeting the specific demands of industrial applications.

MeshGPT (Siddiqui et al., 2023) specializes in creating compact geometries with sharp edges through mesh generation. However, the proposed autoregressive pipeline lacks the ability to support parametric control.

3D-GPT (Sun et al., 2023) introduces a procedural approach, employing LLMs to generate Python code in Blender for 3D scene creation, notably enhancing semantic precision in object modeling. It effectively interprets specific descriptions of a class of objects, producing satisfactory outcomes. However, in the realm of industrial applications, which our research targets, descriptions typically include specific, precise parameters of each object. Moreover, our evaluation emphasizes exact dimensional parameters, in contrast to previous studies that primarily rely on semantic metrics like the CLIP score. Consequently, our work delves into achieving a higher level of precision through parametric control.

Makatura et al. (2023) evaluates the capabilities of GPT-4 in computational design and manufacturing, presenting cases and analysis in shape modeling of LLM using Constructive Solid Geometry. However, the investigation revolves around design abilities given high-level inputs, whereas our work extends to parametric modeling given detailed inputs.

## 2.2 LARGE LANGUAGE MODELS FOR PROGRAM SYNTHESIS

Recent advancements have seen large language models (LLMs) like GPT-4 (OpenAI, 2023) and Code Llama (Rozière et al., 2023) excel in generating code snippets from natural language inputs. These models have demonstrated remarkable proficiency in writing Python functions on benchmarks such as HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021). In response to their rapid development, more comprehensive benchmarks like ClassEval Du et al. (2023), which assesses class-level Python code, and HumanEval-X (Zheng et al., 2023), which extends to multilingual programming in languages like C++, Java, and JavaScript, have been introduced. Despite these advancements, the specific application of LLMs in 3D object modeling remains under-explored. 3D-PreMise is the pioneering benchmark dataset focused on program synthesis for 3D object modeling, delving into the nuances of 3D geometric reasoning within programming. While existing code generation benchmarks evaluate programming logic, 3D-PreMise investigates further on 3D modeling, which involves spatial reasoning and geometric calculation.

Therefore, the potential of LLMs’ capacity in 3D parametric modeling remains largely untapped. To bridge this gap, we develop 3D-PreMise to scrutinize LLMs and generation strategies, with the goal of catalyzing progression in this area.

## 3 PIPELINE

In this section, we introduce a pipeline that integrates LLMs, 3D modeling platforms, and a multimodal interface to facilitate iterative refinement. The methodology unfolds across the following stages, as depicted in Figure 1.

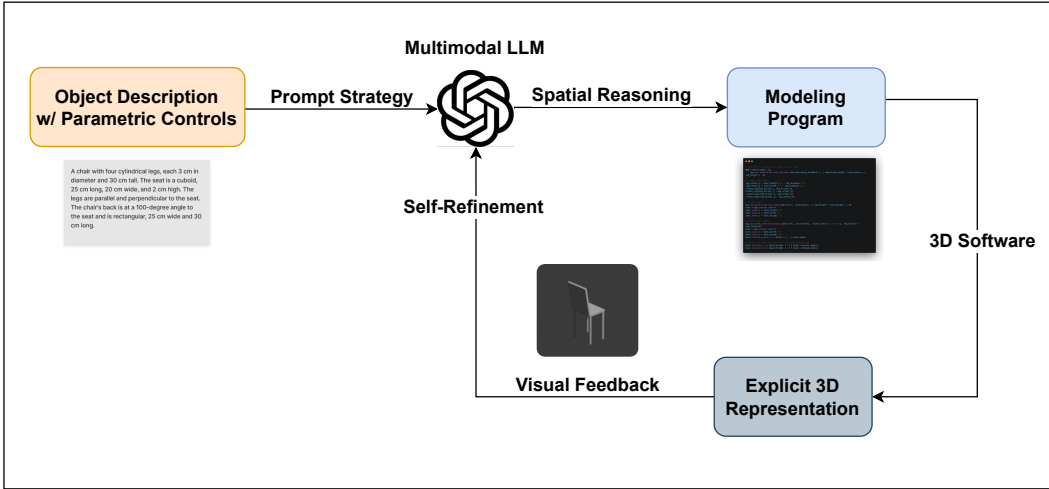


Figure 1: Pipeline Overview

The pipeline takes in text input that precisely defines an object and incorporates strict parametric controls. To augment the LLM’s processing and reasoning capabilities, we employ prompting strategies such as in-context learning and chain-of-thought prompting. The LLM then undertakes spatial reasoning to synthesize the modeling programs. These programs are executed on a 3D modeling platform to yield an explicit 3D representation of the object. Subsequently, we render the object to produce visual feedback. This visual output is then fed back into the multimodal LLM, enabling iterative self-refinement to enhance object modeling.

---

## 4 BENCHMARK DATASET 3D-PREMISE

In this section, we elaborate on our proposed benchmark dataset, 3D-PreMise.

### 4.1 OVERVIEW

Our dataset comprises 57 samples, each consisting of two parts: a prompt and a canonical modeling program. The dataset’s emphasis is on standalone, rigid objects typical in industrial design settings.

In terms of evaluation, we employ a specialized testing program. It analyzes the generated modeling programs and determines the outcome as either ‘pass’ or ‘fail’. This is further quantified by a similarity distance metric that gauges the accuracy of the generated object against the prompt’s specifications.

### 4.2 DATA SAMPLES

A data sample in 3D-PreMise contains specific descriptions of an object and a ground-truth modeling program. As an illustration, a sample is shown as follows.

#### Prompt: Object Description

A chair with four cylindrical legs, each 3 cm in diameter and 30 cm tall. The seat is a cuboid, 25 cm long, 20 cm wide, and 2 cm high. The legs are parallel and perpendicular to the seat. The chair’s back is at a 100-degree angle to the seat and is rectangular, 25 cm wide and 30 cm long.

#### Visualization: Object Shape



Figure 2: Example object description in 3D-PreMise

The prompts in the dataset are meticulously crafted to describe the shape of an object with specific definition detailing parameters.

Correspondingly, the canonical solution presents a correctly formulated Python program using the ‘bpy’ package for Blender that models the object shape as described in the prompt. Each script typically starts with importing necessary libraries, setting up the environment for 3D operations. Variables are declared to hold specific measurements, reflecting the dimensions of the object being modeled, such as lengths, widths, heights, and angles. These variables allow for parametric control over the object’s geometry and are often accompanied by comments that indicate the real-world size they represent. Functions are a key component of the code, designed to encapsulate the creation of different parts of the object. For instance, a function might be dedicated to creating a single element like a leg of a chair or a wheel of a vehicle, with parameters that allow for position and size customization. The main body of the script calls these functions and sets properties to assemble the complete object. It uses transformation operations such as scaling, rotation, and translation to position elements correctly in the 3D space. Lastly, the code is tailored to interact with specific 3D software APIs through Blender’s ‘bpy’ package. This interaction includes the use of functions to add primitives and modify object properties.

### 4.3 CHARACTERISTICS

**Scale and Diversity.** Our dataset consists of 57 samples and encompasses a range of categories, such as furniture, toys, and decorative items.

**Object Geometrical Characteristics.** Our dataset targets objects prevalent in industrial design and manufacturing, characterized by their strictly defined shapes, usually composed of fundamental geometries such as cuboids and cylinders.

**Object Description Characteristics.** The prompt descriptions of objects in our dataset are almost fully defined, leaving little space of ambiguity, which emulates the scenarios in industry applica-

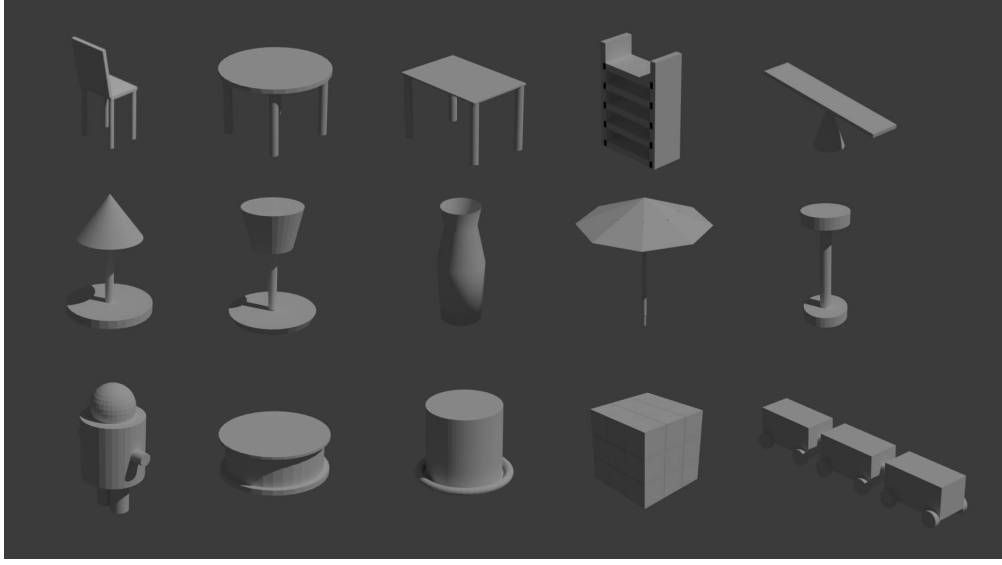


Figure 3: Objects in 3D-PreMise Dataset

tion. In this way, we aim to enable replacing complex engineering drawings with natural language interface for controllable industrial design.

#### 4.4 TEST PROGRAM

While most works on 3D object generation adopt metrics like Fréchet inception distance (FID), CLIP R-Precision or user studies, our setting targets precise parametric modeling. Therefore, we implement a test program to evaluate if the result is accurately faithful to prompt description.

##### 4.4.1 GENERATION CORRECTNESS

The test program first executes the synthesized and ground-truth programs within a virtual sand-box to retrieve the mesh data,  $M_S$  and  $M_T$ , generated using Blender 'bpy' package. To enhance efficiency, we extract only the vertices to form two point clouds,  $P_S$  and  $P_T$ , for testing, which empirically satisfies our demand.

To eliminate the influence of position and orientation on geometric judgement, we normalize the point clouds into  $P'_S$  and  $P'_T$  as follows.

1. Translation to Origin: Calculate the geometric center  $\mathbf{c}$  of each point cloud and translate all points such that the geometric center moves to the origin. The transformation for a point  $\mathbf{p}$  in point cloud  $P$  is given by:

$$\mathbf{p}' = \mathbf{p} - \mathbf{c}$$

where  $\mathbf{c} = \frac{1}{|P|} \sum_{\mathbf{p} \in P} \mathbf{p}$ .

2. Calculation of Principal Axes: Compute the covariance matrix  $\mathbf{C}$  of the translated point cloud. The principal axes are the eigenvectors  $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$  of  $\mathbf{C}$ , corresponding to its eigenvalues. The covariance matrix is defined as:

$$\mathbf{C} = \frac{1}{|P|} \sum_{\mathbf{p}' \in P} (\mathbf{p}' - \bar{\mathbf{p}}')(\mathbf{p}' - \bar{\mathbf{p}}')^T$$

where  $\bar{\mathbf{p}}'$  is the mean of the translated points.

3. Rotation Transformation: Apply a rotation matrix  $\mathbf{R}$  to align the principal axes with the coordinate axes. The rotation matrix is formed by the eigenvectors as its columns:

$$\mathbf{R} = [\mathbf{e}_1 \ \mathbf{e}_2 \ \mathbf{e}_3]$$

Each point  $\mathbf{p}'$  in the translated point cloud is then transformed as:

$$\mathbf{p}'' = \mathbf{R}^T \mathbf{p}'$$

After these transformations, the normalized point clouds  $P'_T$  and  $P'_S$  are obtained, which are independent of their original position and orientation.

Following the normalization, the next step involves comparing the synthesized point cloud  $P_S$  with the ground-truth point cloud  $P_T$  derived from the canonical solution. This process is executed as follows:

1. Point Matching: For each point  $\mathbf{p}'_i$  in the synthesized point cloud (e.g.,  $P'_S$ ), identify the nearest point  $\mathbf{q}'_i$  in the ground-truth point cloud (e.g.,  $P'_T$ ). The distance between these two points is calculated using the Euclidean distance formula:

$$d(\mathbf{p}'_i, \mathbf{q}'_i) = \sqrt{(\mathbf{p}'_i - \mathbf{q}'_i) \cdot (\mathbf{p}'_i - \mathbf{q}'_i)}$$

2. Success Criteria: The match is considered successful if the distance  $d(\mathbf{p}'_i, \mathbf{q}'_i)$  does not exceed a predefined threshold  $\delta$ :

$$\text{Match Success} \Leftrightarrow d(\mathbf{p}'_i, \mathbf{q}'_i) \leq \delta$$

3. Reverse Matching: To ensure the robustness of the match, a reverse matching process is also conducted. In this process, each point in the ground-truth point cloud is matched to the nearest point in the synthesized point cloud, following the same distance criteria.
4. Final Decision: If all the points in both point clouds successfully match with their counterparts under these criteria, the test program outputs a 'pass'. Otherwise, it outputs a 'fail'.

#### 4.4.2 DEVIATION DISTANCE

To further quantify the deviation between the generated code and ground-truth solution, we introduce Chamfer distance on vertex location, defined as follows, where  $V_1$  and  $V_2$  are two point clouds to be compared.

$$d_{CD}(P'_S, P'_T) = \frac{1}{|P'_S|} \sum_{x \in P'_S} \min_{y \in P'_T} \|x - y\|_2^2 + \frac{1}{|P'_T|} \sum_{x \in P'_T} \min_{y \in P'_S} \|x - y\|_2^2$$

## 5 EXPERIMENTS

Using 3D-PreMise, we conduct experiments and analysis with state-of-the-art LLMs on 3D shape generation through modeling program synthesis.

### 5.1 GENERATION STRATEGIES AND PROMPT DESIGN

In this section, we explore the impact of various generation strategies on the performance of LLMs within the scope of 3D-PreMise. Our experiment is structured around the following strategies:

- Zero-Shot Baseline: Directly input the task requirement and description.
- Zero-Shot Chain-of-Thought: Instruct the LLM to answer progressive questions and think step by step before generating code. The questions include listing specific shape, parameters and spatial position of each element.
- One-Shot In-Context Learning: Provide an example containing a description and an answer.
- Few-Shot In-Context Learning: Provide 3 examples containing descriptions and answers.
- One-Shot Chain-of-Thought: Provide an example containing description, reasoning and answer. In addition, instruct the LLM to answer progressive questions before generating code.

## 5.2 IMPLEMENTATION DETAILS

We adopt the OpenAI API for model (OpenAI, 2023) interface in December, 2023. The maximum window size is set to 1024 tokens. As for inference settings, we follow the mainstream strategies in recent code generation works (Du et al., 2023) :

1. Greedy Sampling: setting the temperature as 0 to generate one greedy sample and calculate  $Pass@1$  metric.
2. Nucleus Sampling: setting the temperature as 0.9 to generate 3 samples and calculate  $Pass@k$ , where  $k = 1, 3, 5$ .

## 5.3 QUANTITATIVE RESULTS

The experiment results are quantified using the common metric  $pass@k$  (Kulal et al., 2019):

$$pass@k = \mathbb{E}_{\text{Problems}} \left[ 1 - \frac{\binom{n-c}{k}}{\binom{n}{k}} \right]$$

The results under different settings are shown below.

Table 1: Experiment Results on 3D-PreMise

Model	Generation Strategy	Greedy Sampling	Nucleus Sampling		
		Pass@1	Pass@1	Pass@3	Pass@5
GPT-4	Zero-Shot	7.0%	3.5%	5.3%	7.0%
GPT-4	Zero-Shot CoT	5.3%	5.3%	7.0%	7.0%
GPT-4	One-Shot	12.3%	12.3%	14.0%	14.0%
GPT-4	Few-Shot	12.3%	12.3%	14.0%	15.7%
GPT-4	One-Shot CoT	17.5%	15.8%	19.3%	19.3%

According to the quantitative results, we can summarize the following findings.

- **Finding 1:** In-context learning enhances the LLM performance. However, the performance gap between one-shot and few-shot scenarios is marginal. This observation aligns with expectation, considering the difficulty in deciphering the internal process of analysis between description and modeling code.
- **Finding 2:** The adoption of chain-of-thought prompting notably advances the performance. Nonetheless, the inclusion of an explicit example of the analysis process is essential for this improvement. The provision of guiding questions alone does not yield any measurable enhancement.

## 5.4 CASE ANALYSIS

In this section, we examine the incorrect cases found in our experiments. To better understand these issues, we’ve grouped the errors into several categories.

- **Syntax Error:** issues related to the modeling code, which may prevent correct execution.
- **Geometric Error:**
  - **Structural Configuration Error:** the arrangement of geometric elements fundamentally contradicts the description.
  - **Spatial Precision Error:** minor inaccuracies in spatial parameters, including slight deviations in the position, scale, or rotation of specific elements.
- **Logical Error:** errors that arise from a lack of adherence to real-world logic or common-sense principles.

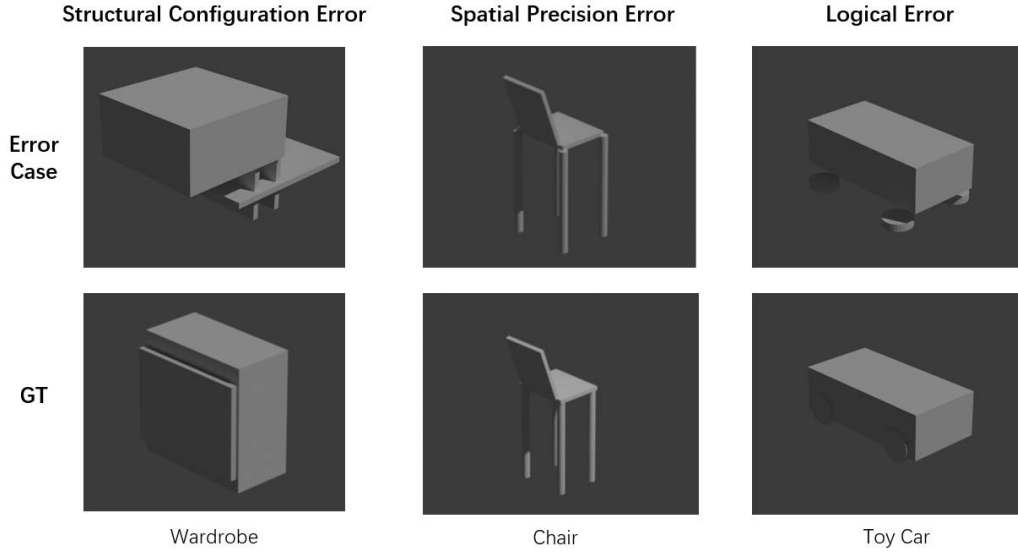


Figure 4: Illustration of Error Categories

To navigate the problem, we conduct a detailed statistical analysis of the frequency in which each error category occurred. The experiment is based on GPT-4 model with temperature set to 0.9.

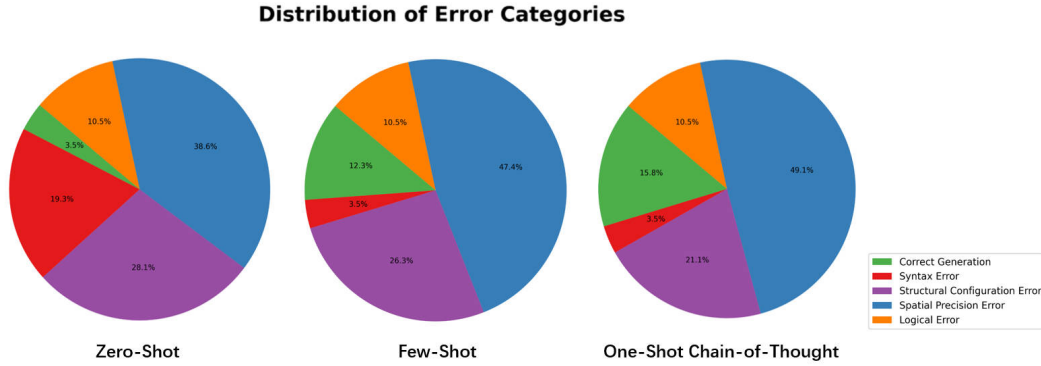


Figure 5: Statistical Analysis on Failure Cases

The analysis of error distributions leads to the following findings.

- **Finding 1:** The generation suffers from spatial precision error most frequently, which is aligned with our expectation that precise control is the painpoint for 3D shape generation.
- **Finding 2:** For cutting-edge LLMs like GPT-4, appropriate in-context learning strategy can significantly alleviate syntax error, making modeling program a feasible representation.
- **Finding 3:** Chain-of-Thought strategy proves to be effective in improving LLM’s structural arrangement when modeling objects.

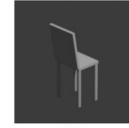




[Task Description]: You are an expert in modeling 3D objects by writing bpy code from Blender in Python. You will be given ...

[In-context Examples]: Here is an example; (Object Description omitted). Analysis of this problem: (Analysis omitted). The program answer to the example is: (Modeling Program omitted)

[Object Description]: Now model this following object: A chair with four cylindrical legs, each 3 cm in diameter and 30 cm tall. The seat is a cuboid, 25 cm long, 20 cm wide, and 2 cm high. The legs are parallel and perpendicular to the seat. The chair's back is at a 100-degree angle to the seat and is rectangular, 25 cm wide and 30 cm long. Please think step by step as shown in the example.



GT



To model the described chair using Blender's Python API, we need to ...  
(Analysis omitted)  
Here's the bpy code for modeling the chair:

```
(only parameter-related part shown)

# Function to create a cuboid
def create_cuboid(name, length, width, height, location):
    bpy.ops.mesh.primitive_cube_add(size=1, location=location)
    cuboid = bpy.context.object
    cuboid.name = name
    cuboid.scale = (length / 2, width / 2, height / 2)
    return cuboid

# Parameters
leg_diameter = 0.03 # 3 cm
leg_height = 0.3 # 30 cm
seat_length = 0.25 # 25 cm
seat_width = 0.20 # 20 cm
seat_height = 0.02 # 2 cm
back_width = seat_length # 25 cm
back_height = 0.30 # 30 cm
back_thickness = 0.02 # Assuming some thickness for the back
back_angle = math.radians(100) # converting degrees to radians

# Creating the seat
seat = create_cuboid("Seat", seat_length, seat_width, seat_height, (0, 0, leg_height))

# Creating the legs
leg_positions = [
    (seat_length / 2 - leg_diameter / 2, seat_width / 2 - leg_diameter / 2, leg_height / 2),
    (-seat_length / 2 + leg_diameter / 2, seat_width / 2 - leg_diameter / 2, leg_height / 2),
    (seat_length / 2 - leg_diameter / 2, -seat_width / 2 + leg_diameter / 2, leg_height / 2),
    (-seat_length / 2 + leg_diameter / 2, -seat_width / 2 + leg_diameter / 2, leg_height / 2)
]
for i, pos in enumerate(leg_positions):
    create_cylinder(f"Leg{i+1}", leg_diameter, leg_height, pos)

# Creating the back
back = create_cuboid("Back", back_thickness, back_width, back_height,
    (0, -seat_width / 2 - back_thickness / 2, leg_height + seat_height + back_height / 2))
back.rotation_euler[0] = math.pi / 2 - back_angle # rotate to 100 degrees from the vertical

(summary omitted)
```



First Generation



Review this rendered image of the object created from your code. Does it meet the requirements of the initial description? If not, identify the specific parameter discrepancies. Please adjust your code based on this analysis.



Visual Feedback



The image shows a chair with its components disassembled and not correctly positioned. The legs and the backrest are not attached to the seat, and the backrest is incorrectly oriented. Here's what needs to be corrected:  
1. Leg Positioning: The legs should be positioned at the corners of the seat, not floating separately.  
2. Backrest Orientation: The backrest is standing upright and not at the 100-degree angle relative to the seat as described.  
3. Assembly: All parts should be joined to form a coherent chair structure.  
Let's correct the code to fix these issues:

```
(only parameter-related part shown)

# Function to create a cube for the seat and back
def create_cube(dimensions, location):
    bpy.ops.mesh.primitive_cube_add(size=2, location=location)
    obj = bpy.context.object
    obj.scale.x = dimensions[0] / 2
    obj.scale.y = dimensions[1] / 2
    obj.scale.z = dimensions[2] / 2
    return obj

# Create the seat
seat_dimensions = (0.25, 0.2, 0.02) # 25cm x 20cm x 2cm
seat_position = (0, 0, 0.3 + 0.01) # 30cm leg height + 1cm seat height above the ground
seat = create_cube(seat_dimensions, seat_position)

# Create the legs
leg_diameter = 0.03 # 3cm
leg_height = 0.3 # 30cm
leg_positions = [
    (seat_dimensions[0] / 2, seat_dimensions[1] / 2, leg_height / 2),
    (-seat_dimensions[0] / 2, seat_dimensions[1] / 2, leg_height / 2),
    (seat_dimensions[0] / 2, -seat_dimensions[1] / 2, leg_height / 2),
    (-seat_dimensions[0] / 2, -seat_dimensions[1] / 2, leg_height / 2)
]

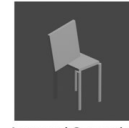
for position in leg_positions:
    create_cylinder(leg_diameter, leg_height, position)

# Create the back
back_dimensions = (0.02, 0.25, 0.3) # 2cm thick, 25cm width, 30cm height
back_position = (0, -seat_dimensions[1] / 2, leg_height + seat_dimensions[2] + back_dimensions[2] / 2)
back = create_cube(back_dimensions, back_position)

# Rotate the back around the X-axis to tilt it backward
back.rotation_mode = 'XYZ'
back.rotation_euler[0] = math.radians(100 - 90) # 100 degrees from the vertical axis, adjusting to Blender's coordinate system

# Reposition the back to ensure it pivots from the bottom edge
back.location.y -= back_dimensions[2] / 2 * math.sin(back.rotation_euler[0])

(summary omitted)
```



Improved Generation

Figure 6: **Self-Improving with Visual Feedback.** Demonstration of GPT-4's ability to self-correct modeling programs via its visual interface. The red line highlights the modification that resolved the issue with the seat's scale.

Further analysis at the program level of failure cases in parametric control has revealed the following key insights:

- **Finding 1:** The most frequent errors stem from incorrect programming practices, notably the redundant scaling applied both within function definitions and function parameters.
- **Finding 2:** Parametric error rates diminish when parameters described in the prompt are explicitly defined as variables in the code. This practice not only enhances the code’s readability but also facilitates more effective reasoning.

## 5.5 EVALUATING SELF-ENHANCEMENT IN MULTIMODAL LLM VIA VISUAL FEEDBACK

In this section, we examine how the cutting-edge multimodal LLM, GPT-4, can self-improve its modeling program from visual feedback. We use the OpenAI ChatGPT interface to interact with the GPT-4 model in December, 2023.

We employ a one-shot chain-of-thought strategy to input the requirements and description. The LLM-generated code is then rendered in Blender, with the resulting image fed back into GPT-4’s visual interface for self-correction of the modeling program. This iterative process is used to scrutinize and refine the generated code.

As illustrated in Figure 6, GPT-4 initially erred in creating the cuboids for the seat and back. It developed a function that inadvertently halved the scale based on the input length and width, leading to a disproportionately small seat that did not align with the chair’s legs. Upon reviewing the visual feedback from the rendered result, GPT-4 modified the function by adjusting the size parameter in the ‘primitive\_cube.add’ function from 1 to 2. While this approach to coding was unconventional, it effectively corrected the initial mistake, demonstrating that the LLM was capable of identifying and addressing the error.

Our expanded experiments yielded several insights regarding the self-correcting abilities of LLMs through multimodal interfaces:

- **Finding 1:** GPT-4 is able to rectify obvious parametric mistakes arising from incorrect programming practices.
- **Finding 2:** GPT-4 shows a lack of sensitivity to common sense errors discernible from visual feedback, such as a cylindrical chair leg incorrectly positioned partly outside, rather than fully beneath, the seat.
- **Finding 3:** The self-correcting capabilities of GPT-4 are notably limited in scenarios involving complex mathematical calculations in geometry. (e.g. they struggle to correct errors stemming from the misapplication of trigonometric functions like sine and cosine.)
- **Finding 4:** GPT-4 hallucinates in response to the visual feedback, sometimes offering contradictory statements.

## 6 CONCLUSION AND FUTURE WORK

In this study, we have developed a pipeline designed for the generation of 3D shapes with parametric controls and engineering semantics, harnessing the power of multimodal LLMs to utilize 3D software through program synthesis. We introduced 3D-PreMise, a comprehensive dataset supported by a specialized testing program, to critically assess the capabilities of LLMs within this innovative context. Our investigation into various generative strategies has identified key techniques that significantly enhance model performance in different dimensions. We have also explored the effectiveness of a visual interface in augmenting the self-correction abilities of LLMs. Our experiments and analysis have revealed the capacities of LLMs in spatial reasoning, geometric computing, program synthesis and multimodal self-correction.

In the future, we plan to enlarge our dataset and apply efficient fine-tuning to improve our pipeline. Specifically, we aim to enhance LLM’s capabilities in spatial computation, geometric commonsense and the capacity for self-refining 3D modeling programs based on visual feedback, emulating human cognition.

---

## ACKNOWLEDGMENTS

We would like to thank Kenan Yu and Jiaying Lai for their contribution to dataset construction and insightful discussion.

## REFERENCES

- B. Mildenhall, Pratul P. Srinivasan, Matthew Tancik, J. Barron, R. Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *European Conference on Computer Vision*, 2020. doi: 10.1007/978-3-030-58452-8\_24.
- Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *arXiv preprint arXiv: 2308.04079*, 2023.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arXiv: 2006.11239*, 2020.
- Chunyi Sun, Junlin Han, Weijian Deng, Xinlong Wang, Zishan Qin, and Stephen Gould. 3d-gpt: Procedural 3d modeling with large language models. *arXiv preprint arXiv: 2310.12945*, 2023.
- Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning acting and planning in language models. *arXiv preprint arXiv: 2310.04406*, 2023.
- N. Khalid, Tianhao Xie, Eugene Belilovsky, and T. Popa. Clip-mesh: Generating textured meshes from text using pretrained image-text models. *ACM SIGGRAPH Conference and Exhibition on Computer Graphics and Interactive Techniques in Asia*, 2022. doi: 10.1145/3550469.3555392.
- Ben Poole, Ajay Jain, J. Barron, and B. Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *International Conference on Learning Representations*, 2022. doi: 10.48550/arXiv.2209.14988.
- Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, S. Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. *Computer Vision and Pattern Recognition*, 2022. doi: 10.1109/CVPR52729.2023.00037.
- Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *ICCV*, 2023.
- Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolific-dreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv: 2305.16213*, 2023.
- Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Ben Mildenhall, Nataniel Ruiz, Shiran Zada, Kfir Aberman, Michael Rubenstein, Jonathan Barron, Yuanzhen Li, and Varun Jampani. Dreambooth3d: Subject-driven text-to-3d generation. *ICCV*, 2023.
- 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*, 2023.
- 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.
- Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv: 2309.16653*, 2023.
- Yawar Siddiqui, Antonio Alliegro, Alexey Artemov, Tatiana Tommasi, Daniele Sirigatti, Vladislav Rosov, Angela Dai, and Matthias Nießner. Meshgpt: Generating triangle meshes with decoder-only transformers. *arXiv preprint arXiv: 2311.15475*, 2023.

- 
- Liane Makatura, Michael Foshey, Bohan Wang, Felix Hähnlein, Pingchuan Ma, Bolei Deng, Megan Tjandrasuwita, Andrew Spielberg, Crystal Elaine Owens, Peter Yichen Chen, Allan Zhao, Amy Zhu, Wil J Norton, Edward Gu, Joshua Jacob, Yifei Li, Adriana Schulz, and Wojciech Matusik. How can large language models help humans in design and manufacturing? *arXiv preprint arXiv: 2307.14377*, 2023.
- OpenAI. Gpt-4 technical report. *arXiv preprint arXiv: 2303.08774*, 2023.
- Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. Code llama: Open foundation models for code. *arXiv preprint arXiv: 2308.12950*, 2023.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *arXiv preprint arXiv: 2107.03374*, 2021.
- Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv: 2108.07732*, 2021.
- Xueying Du, Mingwei Liu, Kaixin Wang, Hanlin Wang, Junwei Liu, Yixuan Chen, Jiayi Feng, Chaofeng Sha, Xin Peng, and Yiling Lou. Classeval: A manually-crafted benchmark for evaluating llms on class-level code generation. *arXiv preprint arXiv: 2308.01861*, 2023.
- Qinkai Zheng, Xiao Xia, Xu Zou, Yuxiao Dong, Shan Wang, Yufei Xue, Zihan Wang, Lei Shen, Andi Wang, Yang Li, Teng Su, Zhilin Yang, and Jie Tang. Codegeex: A pre-trained model for code generation with multilingual evaluations on humaneval-x. *arXiv preprint arXiv: 2303.17568*, 2023.
- Sumith Kulal, Panupong Pasupat, Kartik Chandra, Mina Lee, Oded Padon, Alex Aiken, and Percy Liang. Spoc: Search-based pseudocode to code. 2019.