**Proceedings of the ASME 2023
International Design Engineering Technical Conferences and
Computers and Information in Engineering Conference
IDETC/CIE2023
August 20-23, 2023, Boston, Massachusetts**

IDETC/CIE-2023-115081

PERCEIVED COMPLEXITY OF 3D SHAPES FOR SPATIAL VISUALIZATION TASKS: HUMANS VS GENERATIVE MODELS

Lucy McShane[1], Christian Lopez[1,2]
[1]Computer Science, Lafayette College, Easton, PA
[2]Mechanical Engineering, Lafayette College, Easton, PA

ABSTRACT

The objective of this work is to study the perceived complexity of 3D shapes from a human and a large generative model (i.e., ChatGPT) point of view. This work helps to better understand what makes 3D shapes, which are frequently used in spatial visualization tasks, perceived as complex. It also explores how well ChatGPT can capture the consensus of humans as to what makes shapes perceived as complex. Spatial visualization skills are correlated to success in many STEM fields. To enhance Virtual Reality applications aimed at developing spatial visualization skills, models capable of automatically generating shapes of varying complexities could be used to tailor tasks according to users' skill levels. However, it is important to first understand how humans perceive the complexity of 3D shapes, and how this relates to their performance in spatial visualization tasks. The results of this work indicate that some visual features of 3D shapes, like symmetry, are correlated to their perceived complexity and the performance of individuals on spatial visualization tasks. More importantly, the results show that ChatGPT can generate shapes that are perceived as having different degrees of complexity by humans. The findings support the capabilities of large generative models, like ChatGPT, to capture aspects of human consensus, even in subjective matters such as the perceived complexity of 3D shapes. Hence, these models could potentially be used to automatically generate content, like for VR applications, which are tailored to an individual.

Keywords: ChatGPT, Spatial Visualization, Complexity.

1. INTRODUCTION

Spatial Visualization skills can be described as the ability to rotate, manipulate, twist, or invert 3D objects mentally [1], [2]. This skillset is considered one of the most important skillsets in STEM fields, specifically for engineers who often communicate via graphical means [2], [3]. While studies have indicated that this skillset correlates to students' performance, motivation, confidence, and reasoning, many students in the STEM field do not have sufficient spatial visualization skills when they begin their studies [4]–[6]. Unfortunately, coursework and introductory standard instruction might not be enough to develop these skills [7].

Virtual Reality (VR) technology could help users to seek answers and build their knowledge, especially given its increased use in educational settings [8]. This technology has already produced positive outcomes when used to teach and developed spatial visualization skills [9]. However, it has been shown that individuals have differing levels of expertise and skills which can directly impact their performance in a task [10]. Furthermore, while it has been concluded that individuals perform tasks best when the difficulty matches their skill level, most of the educational applications designed to develop spatial visualization skills only allow users to interact with a restricted set of 3D shapes and tasks [11]–[13]. As a result, students could easily become demotivated after limited interaction with the applications [14].

These limitations could be addressed by automatically generating new content for students. In VR applications made to develop spatial visualization skills, automatically generating new 3D shapes of different levels of complexity in accordance with users' skill levels has the potential to improve students' state of flow and increase motivation. Machine Learning methods have the potential to enhance the understanding regarding the perceived complexity of 3D shapes and generate shapes across a wide range of complexities. For example, large generative models, like GPT-3 and Dall-E [15], [16], could be used to automatically generate 3D shapes. Similarly, Procedural Content Generation (PCG) methods based on Reinforcement Leaning approaches could be leveraged to create new content [16]–[18]. Particularly, 3D shapes of different complexities could be automatically generated given a function that measures shape complexity [13].

However, on one hand, it is still not clear if using large generative models to create new 3D shapes might lend itself to easily tuning the complexity between shapes. Since studies have indicated that users have little understanding of the compatibility of their prompt with the generative model, they might need to try a variety of prompts until responses produce desired outputs [19]. On the other hand, PGC methods based on Reinforcement Leaning approaches might be better suited to fine-tuning the complexity of a shape, but they require a reward function that accurately captures perceived complexity.

Numerous studies have suggested using predetermined mathematical formulas at the pixel level to measure shape complexity [20]. However, these metrics focus on the topological definition of a shape and do not necessarily encompass the human-perceived complexity of 3D shapes necessary to develop spatial visualization skills. A system that summarizes the consensus of individuals regarding the perceived complexity of 3D shapes could provide a better understanding of how humans perceive shape complexity. Large generative models and systems, like GPT-3 and ChatGPT, could potentially identify a consensus of what makes some shapes perceived as more complex since they are trained on large representations of human generated data [21]. Studies suggest that generative models have the capacity to reflect complex patterns among humans in different contexts [22]. The use of systems like Chat-GPT (https://openai.com/blog/chatgpt, based on GPT-3) could be used to reflect a uniform understanding of how humans generally perceive complex 3D shapes, in addition to generating computer programming code that can recreate those shapes.

Thus, in this work, ChatGPT and a series of experiments involving human participants are leveraged to analyze the perceived complexity of 3D shapes from both a human perspective and that of a large generative model (i.e., ChatGPT). The objective is to gain a better understanding of the factors that contribute to the perceived complexity in 3D shapes, which are frequently used in spatial visualization tasks. Additionally, this work aims to investigate the ability of ChatGPT, as a large generative model, to capture the consensus among humans regarding the perceived complexity of shapes. Hence, the research questions that guided this work are as follows:

(RQ1) Are there any visual features in 3D shapes used in spatial visualization tasks that correlate with their perceived complexities?

(RQ2) To what extent can ChatGPT capture the elements of human consensus regarding the perceived complexity of 3D shapes?

2. LITERARY REVIEW

2.1 Virtual Reality and Spatial Visualization

VR is applied in numerous disciplines in addition to gaming entertainment including but not limited to health, education, and sports [23]–[25]. Researchers have found that VR helps enhance feelings of presence and immersion when used in the education setting, leaving a long-lasting impact on students [26]. Moreover, VR could assist students in developing a mental model of what they are learning. This allows the student to cognitively interact with the concepts potentially kindling an increased interest in new material [27], [28].

VR is specifically valuable for spatial visualization skills, a complex skill that requires visual abilities and the formation of mental images. This skillset is studied by numerous fields in science, education, and cognitive psychology given its significance across a wide range of disciplines [29]. Spatial intelligence is described as a skillset that helps us comprehend not only visual-spatial tasks and spatial relations but also gains a better orientation of objects in a space [30]. Often, one's spatial ability has a positive correlation with academic performance, particularly in engineering curriculums [20], [25]. VR technology has been proven to aid the learning and development of these skills due to its ability to visualize 3D objects in a 3D virtual setting and facilitate a "first-person" experience [13].

Despite its engagement capabilities, VR could struggle to create meaningful experiences for students in the long run, particularly for applications designed to develop spatial visualization skills, due to potential novelty effects. In the context of VR applications designed to develop spatial visualization skills, not providing users with 3D shapes that are automatically aligned with their skill level, could serve as an obstacle to overcoming issues of long-term engagement and motivation. Hence, the complexity of the 3D shapes used to develop spatial visualization skills should be tailored to users' skill levels to maximize flow and motivation. However, to achieve this, a better understanding of perceived complexity is necessary.

Many researchers have suggested different ways to measure the complexity of shapes, but perceived complexity is difficult to measure algorithmically because it is a broad concept. This work takes steps to approach this problem. It investigates humans' perception of the complexity of 3D shapes, and how it correlates to their performance in spatial visualization tasks. Moreover, it leverages ChatGPT, a large generative model, to gain a better understanding of human consensus regarding what makes 3D shapes be perceived as complex.

2.2 Measuring Complexity

Studies have found strong correlations between the level of complexity and visual characteristics of shapes, like symmetry, clutter, angular variation, curvature, number of elements, openness, and organization [12]. For instance, a study in [31] was conducted to look at the correlation between the number of elements and symmetry of shapes with their perceived complexities. It was found that shapes with more perceived recognizable elements were deemed more complex by the participants. Furthermore, the results indicate that the shapes perceived as more complex by participants were more asymmetrical.

Similarly, other studies, present in [32], have shown that perceived complexity is correlated to variation in an object's curvature in addition to symmetry and a number of distinct elements. Sharper and unpredicted variations indicated a higher complexity. Studies have also recognized that the average perceived complexity rating for surfaces has a direct positive correlation with the variation of surface curvature [32]. Moreover, studies with children using building blocks found that the number of building blocks used to create shapes had a notable positive correlation with the level of perceived complexity [33].

Recently, [34] explored how the visual features of some of the 3D shapes used in the Purdue Spatial Visualization test [35] correlated to their perceived complexity. Participants were asked to rate the perceived complexities of 3D shapes using images and videos of the shapes, as well as to complete the Purdue Spatial Visualization test. The shapes explored in that study were generated using a system of wedges and voxels, so only shapes with no curvatures from the Purdue test were explored. The results indicate that there was a positive correlation between the number of incomplete voxels and inclined planes in a 3D shape and its perceived complexities. Subsequently, a Machine Learning model indicated that the number of elements, symmetry, and surface variability of a shape are critical components that affect perceived complexity and performance in spatial visualization tasks. While some of the results were not statistically significant, the findings aligned with previous research and highlighted that visual features of 3D shapes might be correlated to their perceived complexity [34].

While it should be recognized that progress has been made in capturing the perceived complexity of 3D shapes, it remains imprecise how humans perceive the complexity of 3D objects in the context of spatial visualization tasks. Hence, this work introduces a series of experiments that ask participants to rate the perceived complexity of 3D shapes with different visual features. Moreover, it analyzes the performance of participants in spatial visualization rotation tasks that used the same set of 3D shapes. This is performed with the goal to leverage this data to better understand the consensus of what makes 3D shapes perceived as complex. With this same goal in mind, this work also leverages ChatGPT, a large generative model, to try to capture the human consensus present in its training data of what makes 3D shapes complex.

2.3 Generative Models and Human Consensus

The third-generation Generative Pre-trained Transformer (GPT-3) could be leveraged to better understand how humans holistically perceive the complexities of 3D shapes. GPT-3 is one of the largest language models created to date, built on 175 billion parameters (i.e., greater than the distance between the Earth and the Sun in kilometers), and trained on 670 gigabytes of text data, with the ability to scale immense amounts of data and produce human-like content [36]. Based upon a prompt provided by the user, GPT-3 could generate collections of programming code, words, and/or other data directly. GPT-3, along with other generative models, could generate responses based on human texts as is trained on a large corpus of public internet data which allows them to synthesize and generalize information from its training data [21]. This can be understood similarly to Google's process of reading a prompt and returning relevant responses [22].

Responses from ChatGPT, a chat system based on GPT-3, to a user prompt represent responses like those of human beings present in its training data. Therefore, this could present limitations as responses might show cultural biases including gender, racial, and religious biases in certain contexts. This is why it is critical to frame prompts thoughtfully and contextually, as the response is reliant on what the user provides [21]. As indicated in [37], ChatGPT does not necessarily produce truthful responses, but, rather, it is a tool that could gather a consensus (i.e., pattern) from its training data set. However, it may provide different answers to the same question, and it cannot account for variations particular to an individual [37].

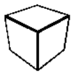
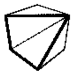
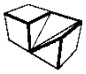
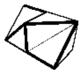
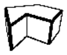
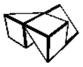
Moreover, large generative models, including GPT-3, have the capability to provide computer programming code alongside textual explanations of the code. A study by [38] was conducted to understand what types of explanations GPT-3 can generate. In the study, 700 code snippets were provided. It showed that GPT-3 could automatically form a checklist of common student mistakes based on the given code snippets, in addition to various explanation types including: (i) fixing bugs with an explanation, (ii) creating analogies to real-world circumstances, (iii) predicting the output in the console, and (iv) listing relevant concepts [38]. This study demonstrates the potential of GPT-3 to provide explanations and summarize commonalities of human performances, and its capability to provide access to quality explanations at a large scale. This is effectively achieved only when the user depicts exactly what they are searching for.

Generative models will only become more sophisticated with their ability to identify patterns that exist among humans in massive amounts of data [21], [22]. Studies provide evidence that algorithmic fidelity, defined as the degree to which the complex patterns within a model accurately reflect ideas, attitudes, and cultures among a range of human subpopulations, is a critical component of generative models [22]. This showcases that generative models could potentially be used in the absence of human data because they could reflect patterns in society present in its training data. Therefore, models that comprise of satisfactory algorithmic fidelity, constitute a powerful tool to enhance the understanding of humans and societies across a wide range of disciplines [22]. Therefore, ChatGPT, as a generative model, could potentially capture consensus across a multitude of topics and disciplines with the capacity to reflect human understandings [21], [36].

Additionally, generative models like GPT3 could be used in relation to Engineering Design. For example, as supported in [39], generative models have the capability to generate responses based on how the user defines the relationship in the hierarchical product structure. Furthermore, generative models can support the synthesis stages of design along with the analysis stages [39]. While some studies have shown the potential of generative models to aid in the Engineering Design Process, others have also indicated that caution needs to be taken since the models might introduce additional biases to the process or even impact high-performing teams [40]–[42].

As it relates to understanding human perceived complexity, while research thus far has explored a variety of ways to measure it, there still lacks a uniform understanding of what generally encompasses human perceived complexity. Thus, in this work, ChatGPT and a series of experiments involving human participants are leveraged to better understand how humans perceive 3D shapes. Moreover, these experiments will help better understand how well ChatGPT, a large generative model, is able to capture some aspects of humans' consensus as to what makes shapes be perceived as complex, which could be

TABLE 1. EXAMPLE OF SHAPES AND THEIR FEATURES

No. of Voxels	No. of complete voxels	No. of incomplete voxels	No. of Incline planes.	Asymmetry metric	3D shape image
1	1	0	0	0	
1	0	1	2	0.19	
2	1	1	3	0.10	
2	0	2	4	0.42	
3	1	2	2	0.14	
3	0	3	3	0.47	

understood as a very subjective topic. This has potential implications for many areas, such as Engineering Design and VR content development.

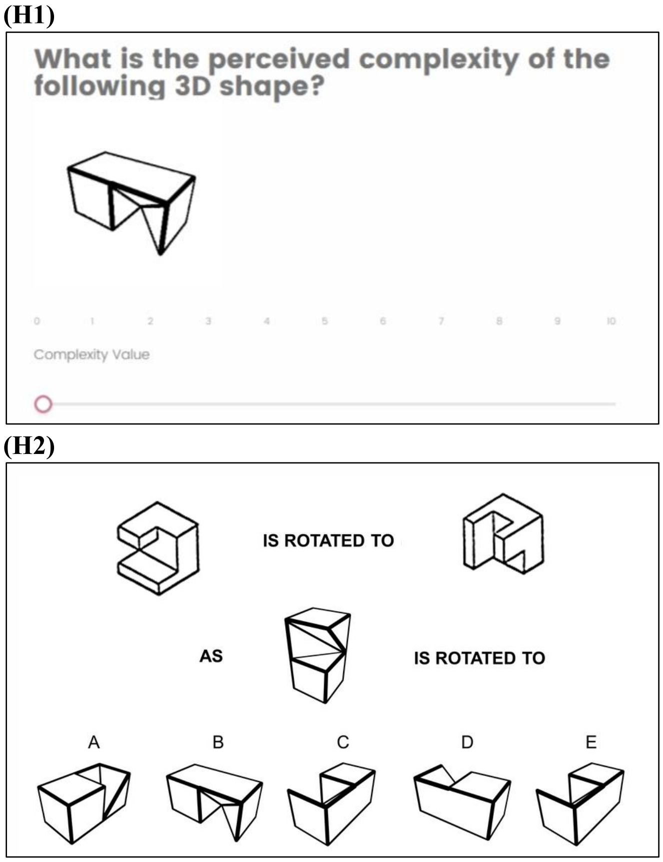
3. METHODS

To address the research questions proposed in this work: first, a set of experiments is conducted with the objective to explore the perceived complexity of 3D shapes, frequently used spatial visualization tasks, with the goal to gain a better understanding of what makes humans perceive certain shapes as more complex. Subsequently, a second set of experiments is conducted to explore ChatGPT's capability in capturing aspects of what makes shapes perceived as complex. The two sets of experiments are introduced next:

3.1. Human Perceived-Complexity Experiments

To explore if there are any visual features in 3D shapes used in spatial visualization tasks that correlate with their perceived complexities, a series of 3D shapes, similar to those present in the Purdue Spatial Visualization test [35], were generated. The same system of voxels and wedges introduced in [34] was used in this study. A total of 39 different shapes were generated with a varying number of voxels (range from 1 to 3 voxels), number of complete voxels (range from 0 to 3 voxels- 0 to 100% of voxels), number of incomplete voxels (range from 0 to 3 voxels- 0 to 100% of voxels), and number of inclined planes (range from 1 to 4 planes) (i.e., independent variables). Moreover, for each of the shapes, an asymmetry metric was calculated based on previous studies (see section 2), and as introduced by [34] (i.e. mediating variable). This asymmetry metric captured how asymmetrical the shapes are in all three axes (see [34] for more details). Table 1 shows multiple examples of the shapes used and their respective visual features and asymmetry metric values. For example, a cube (the first shape on the table) has an asymmetry value of 0 since it is symmetrical in all its axes.

Two sets of questionnaires, H1 and H2, were created to capture the human-perceived complexity of the 3D shapes. H1: In each question, participants were shown an animation of a shape, rotating on all its axes, and were asked to rate the perceived complexity of the shape using a slider bar ranging from 0 (less complex) to 10 (more complex). For this questionnaire, participants were instructed to use a "cube" with a complexity of 0 as a reference shape. This was done to diminish potential individual biases. For this questionnaire, a total of 26 different shapes were shown, that is, 8 shapes of 1 voxel, 8 shapes of 2 voxels, and 8 shapes of 3 voxels respectively. Therefore, the levels of the independent variable of "number of voxels" were assessed within-subject. Of these 26 shapes, two were for quality control (i.e., cube in which they had to select a complexity of 0). The order of the shapes was randomized to avoid any potential order effects.



(H1)

What is the perceived complexity of the following 3D shape?

FIGURE 1. EXAMPLE QUESITONS OF H1 & H2

H2: The second questionnaire was like the Purdue Spatial Visualization test. For this questionnaire participants had to: (i) study how the shape in the top line of the questions is rotated (i.e., reference shape), (ii) picture in their mind what the shape shown in the middle line of the questions (i.e., main shape) looks like when rotated in exactly the same manner, and (iii) select from among the five drawings (A, B, C, D, or E) given in the bottom line of the questions the drawing that looks like the shape rotated in the correct position, which was randomized for each question. Fig. 1 shows an example of the questions present in both the H1 & H2 questionnaires using the same shape.

For all the questions in H2, the same reference shape (i.e., the one in the top line of the questions) and rotation were used to control for any potential effects introduced by these factors. Hence, the only factor that changed between the questions was the main shape. For this questionnaire, a total of 30 shapes were presented, since this was the same number of shapes as in the Purdue test. Of these 30 shapes, two were "cubes" for quality control. No shapes, besides the control, were repeated among participants. However, the shapes were repeated between participants. Also, to closely resemble the Purdue test, participants were given 15 mins to complete it. This time was selected based on the results of some preliminary experiments.

3.2 ChatGPT "Perceived-Complexity" Experiments

ChatGPT was used with the intention of better understanding its capabilities in representing the human perceived complexity of 3D shapes. ChatGPT was prompted with a series of questions asking it to generate code for shapes of different or similar complexities (i.e., the *first set of conversations*), as well as to rank those shapes based on complexity (i.e., a *second set of conversations*). As indicated by previous literature, in generative models like ChatGPT, the prompt provided by the user could impact the model's response according to the context provided. Therefore, to reduce the effects of the context and order of the prompts, for each set of conversations questions were asked randomly.

For the *first set of conversations*, the research team had 17 conversations with ChatGPT. Each conversation began with the prompt "*Create a 3D shape in Python using Plotly"* to ensure it would generate shapes and their respective code using the same Python library. Subsequently, each conversation consisted of 15 questions randomly selected from the following set:

- *Could you create a shape a bit more complex?*
- *Could you create another one of the same complexity?*
- *Could you create a complex shape with Plotly?*
- *Could you create a simple shape with Plotly?*

Hence, the order and quantity of the questions in each conversation were random. From this *first set of conversations*, ChatGPT identified a total of 52 different 3D shapes and provided code to create them using the Python Plotly library. Table 2 shows some examples of the shapes identified. As it can be shown, in some instances ChatGPT provided the names of the shape, while in others it provided a short description as its name. Examples of these shapes, and the ones used in H1 and H2, can be found in the GitHub repo:

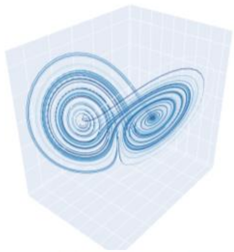
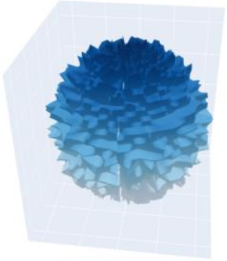
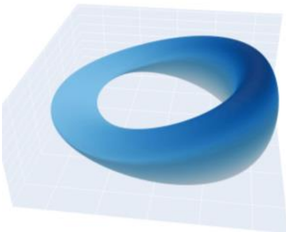
github.com/lopezbec/ShapeComplexity_ChatGPT

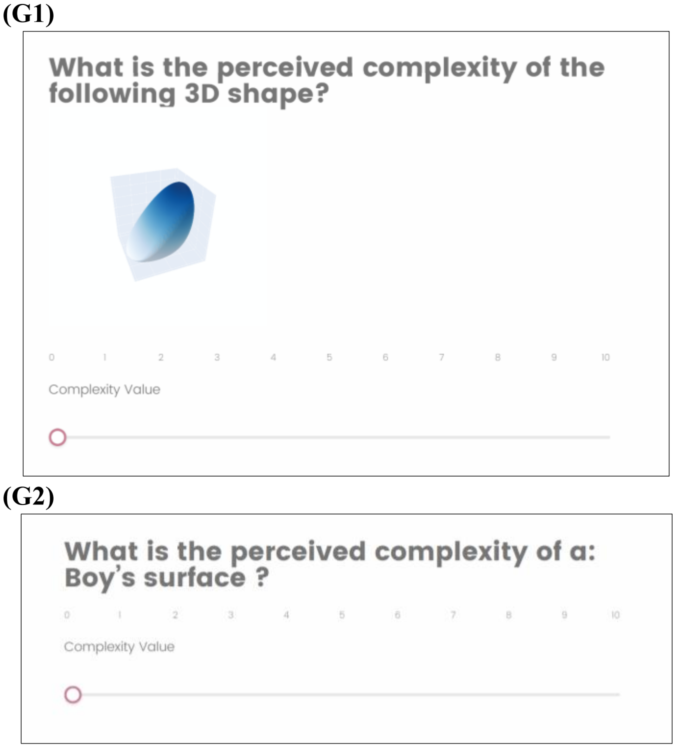
Based on these 52 shapes identified, a *second set of conversations* was performed. For this second set of conversations, the research team had 10 conversations with ChatGPT. Each conversation consisted of 15 questions randomly selected from the following set.

(Q1) *Can you rank the following 3D shapes from least complex to most complex* [list of 52 shapes here]*?*
(Q2) *Which of the following shapes are considered simple shapes* [list of 52 shapes here]*?*
(Q3) *Which of the following shapes are considered complex shapes* [list of 52 shapes here]*?*
(Q4) *Which of the following shapes have similar complexities* [list of 52 shapes here]*?*
(Q5) *Which of the following shapes is the most complex 3D shape* [list of 52 shapes here]*?*
(Q6) *Which of the following shapes is the simplest 3D shape* [list of 52 shapes here]*?*

The questions and the decision of only having 17 and 10 conversations for each set respectively, were based in initial tests that showed consistency in ChatGPT responses. This consistency could be attributed to the use of a "temperature" hyperparameter

TABLE 2. EXAMPLE OF SHAPES FROM CHATGPT

Shapes "Names"	Output of Code
Lorenz Attractor	
Sphere with bumpy surface	
Klein Bottle	



What is the perceived complexity of the following 3D shape?

0 1 2 3 4 5 6 7 8 9 10

Complexity Value

What is the perceived complexity of a: Boy's surface ?

0 1 2 3 4 5 6 7 8 9 10

Complexity Value

FIGURE 2. EXAMPLE QUESTIONS OF G1 & G2

of 0, which controls for the randomness of the text generated by ChatGPT [43].To explore how well ChatGPT could capture the consensus of humans as to what makes shapes be perceived as complex, a second set of questionnaires, G1 and G2, were developed using the 52 shapes identified by ChatGPT in the *first set of conversations*.

G1: In this questionnaire, participants were shown an animation of the shape generated by ChatGPT rotating in all its axes and asked to rate the perceived complexity using a slider bar ranging from 0 (less complex) to 10 (more complex). The research team had to modify some of the code generated by ChatGPT to ensure all the shapes generated had the same format and scale (e.g., color schema, no axis label, no legend, etc.). This was done to avoid introducing any potential bias or confounding factors arising from other visual features of the images (see Table 2).

G2: For this questionnaire, participants were given just the name of the shapes generated by ChatGPT and asked to rate the perceived complexity using a slider bar as well. In this questionnaire, participants were encouraged to search online if they were unfamiliar with the shape. This second questionnaire, which only presented participants with the names of the shapes, was done since it was clear that for some shapes ChatGPT did not generate code that aligned 100% with their names. For example, the Klein Bottle shape in Table 2 does not illustrate a Klein Bottle per se.

For both questionnaires (G1 and G2), a total of 20 different shapes were shown. Two of the questions were randomly placed in the survey for quality control showing a cube in which the participant had to select a complexity of 0. This was done to diminish potential individual biases. Moreover, the order of the shapes was also randomized to avoid any potential order effects. Figure 2 shows an example of a question for both questionnaires (G1) and (G2) using the same shape.

4. RESULTS AND FINDINGS

For this work, participants were recruited via Amazon Mechanical Turk (AMT) [44]. AMT offers low-cost access to a vast and diverse pool of participants, and it has been used largely in behavioral research [45], [46]. A total of 200 participants completed each of the questionnaires (i.e., H1, H2, G1, and G2, see section 3). For each questionnaire, participants were compensated US$0.51 if they completed the questionnaires and correctly answered the two quality control questions. Moreover, for questionnaire H2, participants were compensated an additional US$0.1 for each correct response. This was done with the goal of mimicking the same circumstances of the Purdue test in which participants have the incentive to do well. If the participant did not pass the quality control, they were compensated just US$0.01.

Table 3 shows the number of participants and summary statistics for the completion time of each questionnaire after removing participants that: (i) did not pass the quality control questions, (ii) those that took longer than the 3rd quantile, and (iii) those that took less than the 2nd quantile (e.g., did not read the instructions or just left the survey running).

4.1. Human Perceived-Complexity Experiments Results

For each of the shapes presented in Questionnaire H1, an average perceived complexity value per shape was calculated from all the responses. The average perceived complexity of the 39 different shapes ranged from 1.65 to 7.92 (M: 5.16, Mdn: 5.57, Sd: 1.63), excluding the quality control questions.

A series of correlation tests were performed between the average perceived complexity of the shapes and: (i) their number of voxels, (ii) their number of complete voxels, (iii) their number of incomplete voxels, (iv) their number of incline planes, and (v) their asymmetry metric. Table 4 shows the summary statistics of these tests. These results indicate that on average, the more voxels a shape had, the more incomplete voxels it had, and the more asymmetrical it was, it was perceived as more complex by participants.

For questionnaire H2, each participant was assigned a "grade", representing the proportion of correct responses. This grade ranged from 14.29% to 100% (M: 61.01%, Mdn: 64.29%,

TABLE 3. SUMMARY OF QUESTIONNAIRES PARTICIPAN AND COMPLETION TIME

	No. of participants	*Avg. Completion time [mins]*	*Min. Completion time [mins]*	*Max. Completion time [mins]*
H1	31	4.8	3.8	6.0
H2	48	7.4	4.9	10.7
G1	36	4.6	2.1	7.4
G2	33	4.2	2.7	5.7

TABLE 4. SUMMARY STATISTICS AVG. PERCIVED COMPLEXITY VS SHAPE FEATURES

	ρ	*p-value*
(i) No. of Voxels	0.575	<0.001
(ii) No. of Complete Voxels	-0.088	0.5956
(iii) No. of Incomplete Voxels	0.69	<0.0001
(iv) No. of Incline planes	0.317	0.049
(v) Asymmetry metric	0.444	0.006

Sd: 33.56%), excluding the quality control questions. From this distribution of participants' grades, it is evident that it was negatively skewed suggesting that some participants might have not been motivated by the incentives provided or might have lacked the necessary spatial ability to do well. In either case, the subsequent analyses were performed only using the responses of participants that performed better than average (i.e., a grade of 61.01% or higher).

An average grade per shape was calculated from all the responses of the "above-average" participants. This average grade per shape was statistically significantly correlated with the asymmetry of the shapes (*ρ:*-0.456, *p-value*: 0.006). Similarly, an average completion time per shape was calculated from all the "above-average" participant responses. This average completion time per shape was also statistically significantly correlated with the asymmetry of the shapes (*ρ:*0.501, *p-value*: 0.001). This indicates that the more asymmetrical the shapes on the questions from H2 were, on average participants tended to perform worse. In other words, the more symmetrical the shape was on the spatial visualization tasks, the better participants performed. Moreover, it shows that the more asymmetrical the shapes were, the participants took longer to complete them. This indicates that the "above-average" participants took longer to select their final response in spatial visualization tasks that used shapes that were more complex.

4.2. ChatGPT "Perceived-Complexity" Experiments Results

For each of the shapes presented in questionnaires G1 and G2, an average perceived complexity value per shape was calculated from all the participant responses to each questionnaire. The average perceived complexity of the 52 different shapes, excluding the quality control shape, ranged from 0 to 8.13 (M: 4.76, Mdn: 5.25, Sd: 2.18) based on G1 responses, and ranged from 1.856 to 7.5 (M: 5.29, Mdn: 5.32, Sd: 1.59) based on G2 responses. The average perceived complexity of the shapes based on G1 and G2 were just moderately correlated (*ρ:*0.489, *p-value*:<0.001). This indicates there were some similarities in the responses when participants rated the perceived complexity of the shapes given just the name and when given the shape generated by ChatGPT.

Table 5 shows the summary statistics for the preliminary analyses performed on ChatGPT responses for the *second set of conversations* (see section 3.2). The Kappa values show moderate interrater reliability for ChatGPT's responses to questions Q2, Q3, Q5, and substantial interrater reliability for Q6. However, it shows poor interrater reliability for the responses to questions Q1 and Q4. Moreover, the results show that ChatGPT is more consistent (i.e., greater Kappas values) for questions related to "simple shapes" (i.e., Q2, and Q6) than for "complex shapes" (i.e., Q3, and Q5). Additionally, it shows that ChatGPT is less consistent for questions that require a ranking or a grouping of shapes, which inherently would produce responses of longer length.

TABLE 5. SUMMARY STATISTICS ANALYSIS OF CHATGPT RESPONCES

	Avg. Response length (Mdn)	Kappa (p-value)	G1 vs ChatGPT *ρ* (p-value)	G2 vs ChatGPT *ρ* (p-value
Q1	45.26 (46)	0.12 (<0.0001)	0.071 (0.621)	0.531 (<0.0001)
Q2	12.71 (12)	0.49 (<0.0001)	-0.186 (0.256)	-0.682 (<0.0001)
Q3	17.30 (18.5)	0.4 (<0.0001)	0.233 (0.159)	0.422 (0.008)
Q4	24.59 (19)	0.183 (<0.0001)	0.139 (0.324)	0.363 (0.008)
Q5	4.8 (4)	0.43 (<0.0001)	-0.185 (0.435)	0.316 (0.175)
Q6	1 (1)	0.77 (<0.0001)	-0.389 (0.746)	0.084 (0.947)

Moreover, Table 5, shows the summary statistics for a series of correlation tests performed between the average perceived complexity of the shapes from both G1 and G2 questionnaires, and the proportion of times a given shape was present in ChatGPT's response). Essentially, this analysis looks at the correlation between ChatGPT responses and human perceived complexity response from G1 and G2. This table shows that there are statistically significant correlations between the responses of ChatGPT for questions Q1, Q2, Q3, and Q4, and participants' responses for questionnaire G2. The lack of significant correlation with the responses of participants from questionnaire G1 suggests that there was more agreement with ChatGPT when participants were asked to rate the perceived complexity of a shape based on their name. This could be attributed to the fact that some of the shapes generated by ChatGPT (i.e., the ones used in G1) did not align with their names (e.g., see Table 2).

Finally, a series of t-tests were performed to compare the average complexity of shapes based on responses from questionnaires G1 and G2. The groups compared were formed based on the median response of ChatGPT for questions Q2, Q3, Q5, and Q6 from the *second set of conversations* (see section 3.2). The shapes shown in the responses for Q2, Q3, Q5, and Q6 were transcribed as either present (1) or not present (0) for each response. The median of these results was calculated to create two groups. Subsequently, the average difference in the response from the human participants for G1 and G2 between these two groups was evaluated using a t-test.

TABLE 6. SUMMARY STATISTICS ANALYSIS OF CHATGPT RESPONCES

	G1	G2
	$\mu_{QMdn=1}$ / $\mu_{QMdn=0}$	$\mu_{QMdn=1}$ / $\mu_{QMdn=0}$
	(p-value)	(p-value)
Q2	4.222 / 4.891 (0.292)	3.358/ 5.688 (<0.001)
Q3	5.499/ 4.353 (0.082)	6.081 / 4.726 (0.001)
Q5	5.666 / 4.177 (0.009)	6.117/4.619 (<0.001)
Q6	3.889/4.766 (0.006)	3.313 / 5.232 (<0.0001)

Given that ChatGPT only achieved a moderate agreement on most of the questions responses and some shapes might have only been present a few times over all the responses (i.e., outlier), the median was used to identify the two groups for the t-tests. The median indicates if a given shape was present in more than half of ChatGPT responses, or not, and is also less affected by outliers than the mean. Lastly, the responses for Q1 and Q4 did not lend themselves to dividing the shapes into unique groups since these were ranking questions.

In Table 6 the results from this t-test are shown. The results show that the average perceived complexity of shapes identified by ChatGPT as most complex (i.e., Q5) and simplest (i.e., Q6), were significantly different. This difference was significant even when using both the average perceived complexity estimated from participants' responses to questionnaires G1 and G2. However, for the group generated from ChatGPT's responses to questions Q2 and Q3, only their average perceived complexity estimated from participants' responses to questionnaire G2 (i.e., just showing the names) was statistically significant. This indicates that ChatGPT was in greater agreement with participants' perceived complexity of shapes when presented just with their names.

In summary, the results of this work indicate that:

- The perceived complexity of shapes is positively correlated with (i) their asymmetry, (ii) their number of voxels (e.g., components), and (iii) their number of inclined planes (e.g., surface variability).
- Participants tend to perform worse and take longer in spatial visualization tasks that used asymmetrical shapes (i.e., more complex shapes).
- ChatGPT is more consistent in its response to questions related to less complex (i.e., simple) shapes.
- ChatGPT is less consistent in its response to questions that require ranking or grouping shapes based on complexity.
- ChatGPT is better at generating names of shapes of different complexities than generating code that produces those shapes.

5. CONCLUSION & FUTURE WORKS

Spatial visualization skills are important for STEM fields. Yet, many students do not possess a sufficient skillset in this area. VR technology has been used to help develop spatial visualization skills, but most of the applications are not tailored to the user's skill level, only allowing the student to interact with a limited set of predetermined 3D shapes. Hence, there is potential to leverage Generative Machine Learning methods to generate 3D shapes of different complexities in correspondence with the user's skill level. However, it is important to first understand how humans perceive the complexity of 3D shapes, and how this relates to their performance in spatial visualization tasks.

This work furthers the understanding of what makes 3D shapes perceived as complex, by leveraging ChatGPT and human-participant experiments. The results of this work indicate a positive correlation between the perceived complexity of shapes with their asymmetry, their number of voxels, and their number of incline planes. Furthermore, participants tended to perform worse and take longer in spatial visualization tasks with asymmetrical shapes. Moreover, ChatGPT was proven to show more consistency in its response to questions related to less complex/simple shapes while showing less consistency in questions that required it to group or rank shapes based on their relative complexities. Lastly, ChatGPT performed better in generating names of 3D shapes of varying complexities than generating Python code to create those shapes.

These findings demonstrate that particular features of 3D shapes can help determine the complexity of the shapes and an individual's performance in spatial visualization tasks that use those 3D shapes. These results could support the development of Generative Machine Learning models capable of generating 3D shapes reflective of a desired complexity. This has the potential to help educational applications designed to help develop spatial visualization skills adapt their content and tasks to the users' unique skill level. This ultimately could help further develop these skills that are integral in STEM fields, specifically engineering.

The findings also support the capabilities of large generative models, like ChatGPT, to serve as a tool to reflect patterns in human perceived complexity, which can lead to a better conceptualization of how 3D shapes are perceived as complex. This also supports the potential of these models to identify patterns in their training data that resemble the consensus of humans, which could have implications beyond just generating content. These findings emphasize the potential for using Large Language Models to gather consensus. These models could help assist in identifying effective methods to teach spatial visualization skills contributing to the development of educational applications that are tailored to the needs of the individual.

Nevertheless, this work has several areas for improvement and limitations. One limitation of this study could be attributed to the number of conversations the research team had with ChatGPT. While the decision was informed by the results of initial tests, this was not tested in-depth, and this number could potentially impact the results. Another limitation that could have affected the validity of the results was the distribution of the participants' grades in the questionary H2, suggesting either a lack of motivation or a lack of spatial ability skills. The proportion of responses was negatively skewed, leading to a smaller sample that was used in the analysis (i.e., those who performed "above average"). Additionally, this work illustrated that ChatGPT was not able to consistently produce working Python code for the generation of 3D shapes, indicating that this generative model is not reliable for shape generation per se at this moment. Furthermore, the results show that ChatGPT is less consistent in ranking or grouping shapes of relative complexity. This could be attributed to the broadness and subjectivity of the question, or the size of the list of shapes (i.e., needed to rank 52 shapes). For example, in some of its responses to question Q1, ChatGPT stated that: "*It's difficult to give an absolute ranking of complexity for 3D shapes as different people may have different perspectives on what makes a shape more complex than another. However, based on the general complexity of their shapes and structures, here's an attempt at ranking the provided 3D shapes from least complex to most complex.*" This statement showcases that ChatGPT "recognizes" that while this is a subjective question, there are some general aspect people will have a consensus on.

Future work will look to increase the number of participants, the number of spatial visualization tasks, and the motivation of the participants to obtain a better understanding of perceived complexity. Additionally, future work will aim to tailor the prompt from ChatGPT to be more specific with the intention to decrease the variability in the responses from ChatGPT to gain a narrower and more specific understanding of patterns in human perceived complexities of 3D shapes. This would not only be critical to gain a consensus of perceived complexity to support the development of a VR application to develop students' spatial visualization skills but also to better understand how large generative models could capture key elements of human consensus.

REFERENCES

[1] D. Ben-Chaim, G. Lappan, and R. Houang, "The Effect of Instruction on Spatial Visualization Skills of Middle School Boys and Girls," *American Educational Research Journal*, vol. 25, no. 1, pp. 51–77, 1988.

[2] R. Fleisig, A. Robertson, and A. Spence, "Improving the Spatial Visualization Skills of First Year Engineering Students," *Proceedings of the Canadian Engineering Education Association (CEEA)*, 2011.

[3] S. A. Sorby and B. J. Baartmans, "The development and assessment of a course for enhancing the 3-D spatial visualization skills of first year engineering students," *Journal of Engineering Education*, vol. 89, no. 3, pp. 301–307, 2000.

[4] S. Titus and E. Horsman, "Characterizing and Improving Spatial Visualization Skills," *Journal of Geoscience Education*, vol. 57, no. 4, pp. 229–254, 2009.

[5] M. Omar *et al.*, "Improving Spatial Visualization Skills in Educational Settings," 2022.

[6] B. Verdine, R. Golinkoff, K. Hirsh-Pasek, and N. Newcombe, "Spatial skills, their development, and their links to mathematics.," *Monographs of the society for research in child development*, vol. 82, no. 1, pp. 7–30, 2017.

[7] A. Dayana Farzeeha, M. Omar, and M. Mokhtar, "Spatial visualization ability among engineering students in Malaysia," *Man in India*, vol. 96, no. 1, pp. 203–209, 2016.

[8] E. Hu Au and J. Lee, "Virtual reality in education: a tool for learning in the experience age," *International Journal of Innovation in Education*, vol. 4, no. 4, p. 215, 2017.

[9] G. Makransky, S. Borre-Gude, and R.E. Mayer, "Motivational and cognitive benefits of training in immersive virtual reality based on multiple assessments," *Journal of Computer Assisted Learning*, vol. 35, no. 6, pp. 691–707, 2019.

[10] B. Ottiger *et al.*, "Getting into a 'Flow' state: a systematic review of flow experience in neurological diseases," *Journal of neuroengineering and rehabilitation*, vol. 18, no. 1, pp. 1–21, 2021.

[11] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, "From Game Design Elements to Gamefulness : Defining ' Gamification ,'" *ACM MindTreck'11*, 2011.

[12] C. López and C. Tucker, "Toward personalized adaptive gamification: a machine learning model for predicting performance," *IEEE transactions on Games*, vol. 12, no. 2, pp. 155–168, 2018.

[13] C. Roca-González J. Martin-Gutierrez, M. García-Dominguez, & M. Mato Carrodeguas, "Virtual technologies to develop visual-spatial ability in engineering students," *Eurasia Journal of Mathematics, Science and Technology Education*, vol. 13, no. 2, pp. 441–468, 2017.

[14] M. Virvou and G. Katsionis, "On the usability and likeability of virtual reality games for education: The case of VR-ENGAGE," *Computers & Education*, vol. 50, no. 1, pp. 154–178, 2008.

[16] A. Ramesh, P. Mikhail, G. Gog, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-Shot Text-to-Image Generation." In *International Conference on Machine Learning*, pp. 8821-8831. PMLE, 2021.

[16] C. E. Lopez, O. Ashour, and C. S. Tucker, "Reinforcement learning content generation for virtual reality applications," in *Proceedings of the ASME Design Engineering Technical Conference*, 2019. doi: 10.1115/DETC2019-97711.

[17] J. Cunningham, C. Lopez, O. Ashour, and C. S. Tucker, "Multi-Context Generation in Virtual Reality Environments using Deep Reinforcement Learning," in *Proceedings of the ASME 2020 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE)*, 2020, pp. 1–11.

[18] C. E. López, J. Cunningham, O. Ashour, and C. S. Tucker, "Deep reinforcement learning for procedural content generation of 3D virtual environments," *Journal of Computing and Information Science in Engineering*, 2020, doi: 10.1115/1.4046293.

[19] Y. Zhou *et al.*, "Large Language Models Are Human-Level Prompt Engineers," Nov. 2022, doi: 10.48550/arxiv.2211.01910.

[20] W. Saleem, A. Belyaey, D. Wang, and H. Seidel, "On visual complexity of 3D shapes," *Computers & Graphics*, vol. 35, no. 3, pp. 580–585, 2011.

[21] L. Floridi and M. Chiriatti, "GPT-3: Its Nature, Scope, Limits, and Consequences," *Minds and Machines*, vol. 30, no. 4, pp. 681–694, Dec. 2020, doi: 10.1007/S11023-020-09548-1.

[23] L. Argyle, E. Busby, N. Fulda, J. Gubler, C. Rytting, and D. Wingate, "Out of one, many: Using language models to simulate human samples," *Political Analysis*, 2023, 1-15. doi:10.1017/pan.2023.2

[23] S. Bialkova and M. V. Gisbergen, "When sound modulates vision: VR applications for art and entertainment," *2017 IEEE 3rd Workshop on Everyday Virtual Reality (WEVR)*, 2017.

[24] O. Farley, K. Spencer, and L. Baudinet, "Virtual reality in sports coaching, skill acquisition, and application to surfing," *A review. Journal of Human Sport and Exercise*, vol. 15, no. 3, 2019.

[25] Y. Han, "A Virtual Reality Algorithm for the Study of Clinical Efficacy of Sports Injury Rehabilitation Training," *Journal of Healthcare Engineering*, pp. 1–6, 2021.

[26] M. Vesisenaho *et al.*, "Virtual Reality in Education: Focus on the Role of Emotions and Physiological Reactivity," *Journal For Virtual Worlds Research*, vol. 12, no. 1, 2019.

[27] M. Humphreys, "Developing an educational framework for the teaching of simulation within nurse education.," *Open Journal of Nursing*, vol. 3, no. 04, pp. 363–371, 2013.

[28] W. Alhalabi, "Virtual reality systems enhance students achievements in engineering education," *Behaviour & Information Technology*, vol. 35, no. 11, pp. 919–925, 2016.

[30] Z. Hawes, K. Gilligan-Lee, K. Mix "Effects of spatial training on mathematics performance: A meta-analysis." *Developmental Psychology*, 58(1), pp. 112-137, 2008.

[30] C. Diezmann and J. Watters, "Identifying and Supporting Spatial Intelligence in Young Children," *Contemporary Issues in Early Childhood*, vol. 1, no. 3, pp. 299–313, 2000.

[31] S. Sukumar, D. Page, A. Koschan, and M. Abidi, "Towards understanding what makes 3D objects appear simple or complex," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2008.

[32] I. Biederman and P. Gerhardstein, "Recognizing depth-rotated objects: Evidence and conditions for three-dimensional viewpoint invariance," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 19, no. 6, pp. 1162–1182, 1993.

[34] C. Chen, T. Yang, X. Zhang, and X. Xu,, "Spatial Skills Associated With Block-Building Complexity in Preschoolers.," *Frontiers in Psychology*, vol. 11, 2020.

[34] A. Busheska and C. Lopez, "Exploring the Perceived Complexity of 3d Shapes: Towards a Spatial Visualization VR Application," *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, vol. 86212, no. American Society of Mechanical Engineers, Aug. 2022.

[35] Y. Maeda, S. Y. Yoon, K. Kim-Kang, and P. K. Imbrie, "Psychometric properties of the Revised PSVT:R for measuring First Year engineering students spatial ability," *International Journal of Engineering Education*, vol. 29, pp. 763–776, 2013.

[36] A. Tamkin, M. Brundage, J. Clark †3, and D. Ganguli, "Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models," Feb. 2021, doi: 10.48550/arxiv.2102.02503.

[38] R. Dale, "GPT-3: What's it good for?," *National Language Engineering*, vol. 27, pp. 113–118, 2021, doi: 10.1017/S1351324920000601.

[38] S. MacNeil, A. Tran, D. Mogil, S. Bernstein, E. Ross, and Z. Huang, "Generating Diverse Code Explanations using the GPT-3 Large Language Model," *ICER 2022 - Proceedings of the 2022 ACM Conference on International Computing Education Research*, vol. 2, pp. 37–39, Aug. 2022, doi: 10.1145/3501709.3544280.

[40] O. Isaksson, "A generative modeling approach to engineering design," In *DS 31: Proceedings of Iced 03, the 14th International Conference on Engineering Design, Stockholm*. 2003.

[40] C. E. Lopez, S. R. Miller, and C. S. Tucker, "Exploring biases between human and machine generated designs," *Journal of Mechanical Design, Transactions of the ASME*, vol. 141, no. 2, 2019, doi: 10.1115/1.4041857.

[41] C. E. Lopez, S. Miller, and C. S. Tucker, "Human validation of computer vs human generated desing sketches," in *Proc. ASME Int. Des. Eng. Tech. Conf. Comput. Inf. Eng. Conf.*, 2018, p. 85698.

[43] G. Zhang, A. Raina, J. Cagan, and C. McComb, "A cautionary tale about the impact of AI on human design teams," Design Studies 72, no. 100990, 2021.

[43] Peng *et al.*, "Towards making the most of chatgpt for machine translation," *arXiv preprint arXiv:2303.13780*, 2023.

[44] "Amazon Mechanical Turk." https://www.mturk.com/ (accessed Mar. 10, 2023).

[45] W. Mason and S. Suri, "Conducting behavioral research on Amazons Mechanical Turk," *Behav. Res. Methods*, vol. 44, no. 1, pp. 1–23, 2012.

[46] H. Aguinis, I. Villamor, and R. S. Ramani, "MTurk research: Review and recommendations," *Journal of Management*, vol. 47, no. 4, pp. 823–837, 2021.