

# Understanding Depth and Height Perception in Large Visual-Language Models

## Supplementary Material

The supplementary will provide additional results and analysis on our proposed datasets. Additional results for GeoMeter 2D and GeoMeter 3D datasets are in Section 7.1 and Section 7.2. Sections 8, 9 respectively contain the broader impact and computational resources needed for our work.

### 7. Additional Results

#### 7.1. Quantitative Evaluation

Table 3, Table 4 present detailed results for the GeoMeter 2D dataset; and Table 5, Table 6 present detailed results for the GeoMeter 3D dataset. All of these results examine the impact of scene complexity (3 shapes vs 5 shapes), query attributes (color, labels), and question types (MCQ and True/False) on depth and height perception (respectively). While the main paper reports average results, the individual category-specific outcomes offer deeper insights. For instance, performance deteriorates with increased scene complexity (5 shapes) for many open-source models, highlighting the superior robustness of closed-source models under these conditions. Additionally, changes in query attributes show minimal impact on performance for most models, indicating their resilience to variations in query types.

#### 7.2. Qualitative Examples

Figure 10 displays sample predictions from both open and closed models, highlighting their challenges with depth and height perception. The examples particularly emphasize the models’ inaccuracies, especially in height perception, showcasing their limitations in spatial understanding. This figure includes predictions from the best-performing models in the open (LLaVA 1.5 7B) and closed (GPT 4o) categories. Figures 11 and 12 present examples from the GeoMeter-2D dataset, including the specific prompts for both MCQ and True/False questions, serving as visual aids for the evaluations discussed. Similarly, Figures 13 and 14, showcase samples and corresponding prompts from the GeoMeter-3D depth and height category, respectively. These figures provide insights into the different scenarios and questions used to assess depth and height perception across various data types. Additionally, Figure 15 features image-text pairs from the GeoMeter-2D Basic dataset, highlighting the initial stages of evaluating the models’ capabilities in recognizing basic properties.

### 8. Broader Impact

To our understanding, there are no negative societal impacts of our work. The goal of this work was to evalu-

ate the depth and height perception capabilities of models that may later be used in real-world settings. This research provides insights into the depth and height perception capabilities of vision language models (VLMs), significantly impacting practical applications like autonomous driving, augmented reality, and assistive technologies. This work not only advances theoretical understanding but also opens up new possibilities for real-world applications.

### 9. Computational Resources

All experiments were run on an internal cluster. Each run used a single NVIDIA GPU, with memory ranging from 16GB-24GB.

Table 3. **Performance of the studied models on proposed GeoMeter-2D depth category.** Evaluation is done on the VQA task on MCQ and True/False type questions. Color, RL, PL are the query attributes. Here, RL, PL respectively denotes random numeric label, patterned numeric label.

	Model	Depth-3 shapes						Depth-5 shapes					
		MCQ			T/F			MCQ			T/F		
		Color	RL	PL	Color	RL	PL	Color	RL	PL	Color	RL	PL
Open	LLaVA 1.5 7B	48.0	37.5	54.5	49.0	54.5	47.0	36.5	31.0	39.0	45.0	56.0	49.5
	LLaVA 1.5 13B	36.5	21.0	29.0	52.0	57.0	54.0	35.5	15.0	11.0	54.5	53.0	54.0
	LLaVA 1.6 Mistral 7B	44.0	34.5	25.0	55.5	54.5	52.5	28.5	24.0	11.0	54.0	56.0	54.0
	LLaVA 1.6 Vicuna 7B	37.0	20.5	13.0	54.5	50.5	49.5	29.0	7.0	1.0	50.5	52.5	55.0
	LLaVA 1.6 Vicuna 13B	35.0	42.0	62.0	45.5	53.5	72.0	28.0	35.5	32.0	56.0	54.0	62.5
	Bunny-v1.0-3B	41.5	40.5	38.5	48.0	45.5	54.0	31.0	30.0	13.5	46.5	52.5	55.0
	Bunny-v1.0-4B	38.0	47.0	33.5	55.5	55.5	55.5	26.5	29.5	22.5	52.5	53.0	53.0
	Bunny-v1.1-4B	45.5	47.5	33.5	52.5	55.5	55.5	34.0	36.0	31.5	52.5	53.0	53.0
	Bunny-Llama-3-8B-V	34.5	45.0	46.0	41.0	58.5	51.5	27.5	36.5	48.0	48.5	53.5	46.0
	Fuyu-8B	33.5	17.0	4.5	58.5	55.5	55.5	30.0	15.5	3.0	53.5	53.0	53.0
	InstructBLIP-Flan-T5-XL	45.5	8.5	0.0	44.5	44.5	44.5	32.0	40.0	0.0	47.0	47.0	47.0
	InstructBLIP-Vicuna-7B	43.5	40.0	59.0	49.5	44.0	43.0	32.0	31.0	34.0	46.5	47.5	46.0
	LLaMA-Adapter-v2-Multimodal	41.0	40.0	39.5	48.5	45.5	45.5	31.0	30.0	33.0	47	45.5	45.5
	MiniGPT-4	42.0	41.5	43.0	52.0	51.5	51.5	34.0	32.0	30.0	48.5	47.5	47.5
Closed	GPT-4V	45.0	49.0	41.5	54.5	57.0	61.5	38.5	37.0	40.5	56.0	58.5	53.0
	GPT-4o	47.5	44.5	47.0	55.5	58.5	70.5	49.5	36.5	36.0	62.0	59.0	52.0
	Claude 3 Opus	47.5	40.5	50	51.5	51.5	56.5	36.5	36.0	41.0	52.5	51.5	56.0

Table 4. **Performance of the studied models on proposed GeoMeter-2D height category.** Evaluation is done on the VQA task on MCQ and True/False type questions. Color, Label are the query attributes. Here, SP,  $\overline{SP}$  respectively denote w/ step, and w/o step.

	Model	Height-3 towers $\overline{SP}$				Height-3 towers SP			
		MCQ		T/F		MCQ		T/F	
		Color	Label	Color	Label	Color	Label	Color	Label
Open	LLaVA 1.5 7B	15.5	18.0	50.0	54.0	21.0	16.5	49.5	57.0
	LLaVA 1.5	15.5	9.0	49.0	54.0	14.5	10.0	49.0	56.9
	LLaVA 1.6 Mistral 7B	16.0	17.0	50.5	55.5	14.0	15.5	49.5	53.0
	LLaVA 1.6 Vicuna 7B	14.0	19.0	49.0	55.0	18.5	18.0	50.0	58.0
	LLaVA 1.6 Vicuna 13B	19.0	19.0	49.5	54.0	13.5	20.5	49.5	57.0
	Bunny-v1.0-3B	13.5	17.5	49.0	51.0	18.5	20.0	49.0	57.0
	Bunny-v1.0-4B	18.0	16.5	49.0	54.0	16.0	12.5	49.0	57.0
	Bunny-v1.1-4B	11.0	18.5	49.0	54.0	19.0	15.0	49.0	57.0
	Bunny-Llama-3-8B-V	15.0	15.5	49.0	54.5	14.5	18.0	49.0	53.5
	Fuyu-8B	0.0	0.0	45.5	55.0	0.0	0.0	53.5	55.0
	InstructBLIP-Flan-T5-XL	0.5	0.5	51.0	46.0	0.0	0.5	51.0	43.0
	InstructBLIP-Vicuna-7B	19.0	16.0	52.0	54.0	21.0	20.5	52.5	57.0
	LLaMA-Adapter-v2-Multimodal	11.0	9.0	52.0	50.0	13.0	10.0	53.0	50.0
	MiniGPT-4	13.0	12.0	54.0	52.5	15.0	14.0	54.0	51.5
Closed	GPT-4V	6.5	7.0	48.0	55.5	3.0	10.0	48.5	56.0
	GPT-4o	21.0	17.0	57.0	53.0	17.5	15.5	51.5	56.5
	Claude 3 Opus	15.0	13.5	50.5	51.5	16.0	18.5	50.0	56.0
	Model	Height-5 towers $\overline{SP}$				Height-5 towers SP			
		MCQ		T/F		MCQ		T/F	
		Color	Label	Color	Label	Color	Label	Color	Label
Open	LLaVA 1.5 7B	14.0	14.0	46.0	47.0	14.0	18.5	51.5	51.0
	LLaVA 1.5 13B	12.0	9.0	52.0	49.0	8.5.0	8.0	49.0	48.0
	LLaVA 1.6 Mistral 7B	16.0	14.5	46.0	46.0	17.5	20.5	48.0	51.0
	LLaVA 1.6 Vicuna 7B	16.0	13.5	51.5	49.5	16.0	15.0	48.5	49.0
	LLaVA 1.6 Vicuna 13B	16.5	16.0	52.0	49.0	20.0	14.5	49.0	49.0
	Bunny-v1.0-3B	13.0	11.5	50.5	44.0	12.5	19.5	49.0	50.5
	Bunny-v1.0-4B	16.0	14.5	52.0	49.0	14.0	17.0	49.0	49.0
	Bunny-v1.1-4B	14.5	13.0	52.0	49.0	12.0	18.0	49.0	49.0
	Bunny-Llama-3-8B-V	15.0	15.0	52.0	47.5	14.5	21.0	49.0	49.5
	Fuyu-8B	0.0	0.0	52.5	51.5	0.0	0.0	49.0	46.5
	InstructBLIP-Flan-T5-XL	0.0	1.5	48.0	51.0	0.0	1.5	51.0	51.0
	InstructBLIP-Vicuna-7B	15.0	11.0	52.5	49.0	15.0	16.0	48.5	49.0
	LLaMA-Adapter-v2-Multimodal	10.5	8.5	51.0	52	9.5	9.0	50.0	51.5
	MiniGPT-4	13.5	10.0	52.0	50.0	12.0	10.5	51.0	49.5
Closed	GPT-4V	17.5	12.5	51.5	50.0	14.0	6.5	50.0	49.0
	GPT-4o	18.0	18.5	59.5	50.0	19.0	19.0	51.0	52.0
	Claude 3 Opus	19.5	14.0	48.5	51.5	13.0	19.5	47.5	48.5

Table 5. **Performance of the studied models on proposed GeoMeter-3D height category.** Evaluation is done on the VQA task on MCQ and True/False type questions. Color, ColMat are the query attributes. Here, ColMat denotes color+material

	Model	Depth-3 shapes				Depth-5 shapes			
		MCQ		T/F		MCQ		T/F	
		Color	ColMat	Color	ColMat	Color	ColMat	Color	ColMat
Open	LLaVA 1.5 7B	49.1	42.5	59.4	53.8	43.1	37.5	55.7	50.4
	LLaVA 1.5 13B	51.3	45.9	61.9	58.4	37.3	35.1	50.3	44.3
	LLaVA 1.6 Mistral 7B	47.1	45.3	51.9	50.6	34.8	30.8	50.3	48.9
	LLaVA 1.6 Vicuna 7B	48.8	47.3	61.9	58.3	40.2	32.9	45.9	40.2
	LLaVA 1.6 Vicuna 13B	51.8	50.3	64.2	61.2	48.3	42.9	50.2	45.9
	Bunny-v1.0-3B	34.8	29.3	40.2	35.8	21.9	18.3	34.8	29.8
	Bunny-v1.0-4B	34.2	30.8	45.3	43.2	28.2	23.2	34.9	30.7
	Bunny-v1.1-4B	45.2	40.3	44.2	42.9	40.2	38.3	48.3	42.9
	Bunny-Llama-3-8B-V	44.2	42.1	45.2	40.8	40.8	35.9	40.8	38.3
	Fuyu-8B	41.8	38.4	59.3	51.8	30.5	27.5	48.3	47.2
	InstructBLIP-Flan-T5-XL	58.3	54.2	55.3	51.3	61.9	59.3	54.9	53.8
	InstructBLIP-Vicuna-7B	57.4	56.3	56.9	55.4	60.2	57.3	59.9	58.6
	LLaMA-Adapter-v2-Multimodal	52.9	48.3	47.3	44.2	59.8	56.8	57.8	54.7
	MiniGPT-4	60.3	56.3	57.8	54.8	65.3	62.9	60.3	54.8
Closed	GPT-4V	54.3	50.1	63.9	60.2	45.3	40.9	48.4	43.2
	GPT-4o	59.9	52.9	65.9	60.3	50.3	44.3	50.3	44.8
	Claude 3 Opus	56.3	53.9	57.3	52.3	47.3	43.2	51.8	47.4



Table 6. **Performance of the studied models on proposed GeoMeter-3D height category.** Evaluation is done on the VQA task on MCQ and True/False type questions. Color, ColMat are the query attributes. Here, ColMat, SP,  $\bar{S}\bar{P}$  respectively denotes color+material, w/ step, and w/o step.

	Model	Height-3 towers SP				Height-3 towers SP			
		MCQ		T/F		MCQ		T/F	
		Color	ColMat	Color	ColMat	Color	ColMat	Color	ColMat
Open	LLaVA 1.5 7B	20.3	12.9	48.2	40.8	18.8	8.1	46.3	40.3
	LLaVA 1.5 13B	22.8	18.3	52.1	48.9	19.9	15.8	48.2	45.9
	LLaVA 1.6 Mistral 7B	21.9	18.7	49.9	42.7	18.3	12.8	47.9	44.3
	LLaVA 1.6 Vicuna 7B	20.8	18.9	48.7	44.8	18.7	12.7	49.7	43.8
	LLaVA 1.6 Vicuna 13B	24.9	19.8	50.7	47.3	20.8	17.3	50.2	45.9
	Bunny-v1.0-3B	12.4	9.4	51.4	50.4	9.4	5.3	42.9	40.3
	Bunny-v1.0-4B	14.9	10.4	51.8	48.3	12.9	10.5	44.3	41.7
	Bunny-v1.1-4B	15.9	12.7	54.8	52.6	13.7	11.8	50.3	48.5
	Bunny-Llama-3-8B-V	16.3	12.8	55.7	53.9	14.9	13.9	52.9	49.3
	Fuyu-8B	9.3	7.9	40.2	35.4	5.9	3.9	37.9	34.7
	InstructBLIP-Flan-T5-XL	25.1	20.9	53.8	50.3	22.9	20.4	50.3	48.2
	InstructBLIP-Vicuna-7B	24.9	21.9	54.3	52.9	20.8	18.9	52.7	49.3
	LLaMA-Adapter-v2-Multimodal	23.9	20.3	49.3	47.8	20.2	18.7	48.2	45.8
	MiniGPT-4	26.9	24.8	54.8	53.7	24.8	20.4	53.8	51.8
Closed	GPT-4V	28.8	25.9	48.3	48.0	27.1	26.9	46.0	43.9
	GPT-4o	30.5	28.9	50.9	49.2	28.9	27.8	49.3	46.8
	Claude 3 Opus	28.3	24.0	51.8	48.3	26.1	22.0	47.3	43.0
	Model	Height-5 towers SP				Height-5 towers SP			
		MCQ		T/F		MCQ		T/F	
		Color	ColMat	Color	ColMat	Color	ColMat	Color	ColMat
Open	LLaVA 1.5 7B	12.9	10.4	48.3	42.3	10.4	9.3	47.3	43.8
	LLaVA 1.5 13B	13.9	11.3	50.3	49.2	11.8	10.5	49.3	47.3
	LLaVA 1.6 Mistral 7B	11.0	9.3	50.4	47.3	10.3	8.3	47.0	46.9
	LLaVA 1.6 Vicuna 7B	13.9	10.3	51.9	49.2	11.8	10.8	50.8	47.1
	LLaVA 1.6 Vicuna 13B	15.9	12.3	54.1	50.3	12.9	9.3	52.9	48.3
	Bunny-v1.0-3B	9.2	4.2	34.3	28.4	7.3	6.9	33.2	30.9
	Bunny-v1.0-4B	11.9	9.3	35.3	30.4	9.3	5.3	34.3	33.9
	Bunny-v1.1-4B	13.9	11.4	39.3	36.3	12.9	10.2	37.3	33.9
	Bunny-Llama-3-8B-V	13.3	12.1	38.3	37.9	10.3	9.9	36.3	35.9
	Fuyu-8B	4.2	1.8	35.3	30.0	0.0	0.0	32.8	31.9
	InstructBLIP-Flan-T5-XL	19.8	18.9	47.2	42.1	16.3	15.9	42.9	38.3
	InstructBLIP-Vicuna-7B	18.3	17.9	46.3	45.8	17.0	16.9	43.9	42.7
	LLaMA-Adapter-v2-Multimodal	15.3	12.8	48.3	48.0	13.9	12.8	47.4	45.4
	MiniGPT-4	20.8	19.3	53.2	50.2	19.2	16.0	49.3	47.3
Closed	GPT-4V	19.3	17.3	48.4	47.8	18.3	16.9	47.0	46.3
	GPT-4o	22.6	21.9	51.9	50.3	20.9	19.6	49.4	47.4
	Claude 3 Opus	21.9	19.3	49.3	47.0	19.7	15.9	48.9	44.8

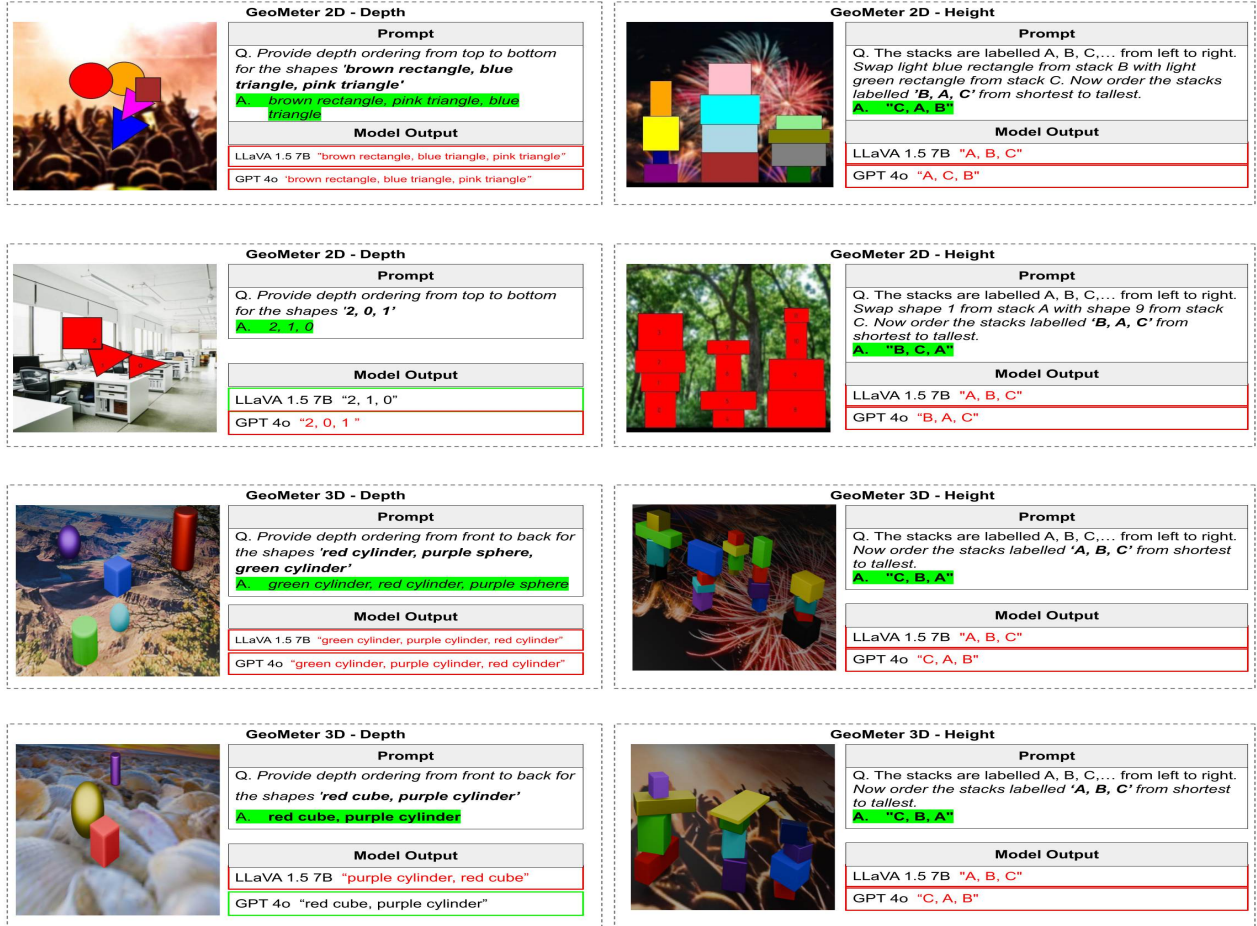
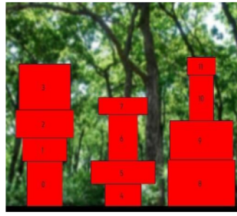


Figure 10. **Depth and height perception of open and closed models.** Here we show the prediction of LLaVA 1.5 7B and GPT 4o. Here Q and A respectively denote Question and Ground Truth Answer. Green and Red boxes respectively denote correct and incorrect prediction.

	<p><b>Prompt</b></p> <p>The image shows 2D shapes placed randomly. The shapes overlap each other, creating a depth effect. When two shapes are overlapping, the shape that is complete is defined to be on top of the partially hidden shape. Each 2D shape has a number written over them which we call ShapeID and must be inferred as the label for the corresponding shape. Provide depth ordering from top to bottom for the shapes '2, 0, 1' in the image. Answer in the format: 'ShapeID, ShapeID, ...'. For eg. '3, 1, 2' is a valid answer format.</p>	<p><b>MCQ</b></p> <p>Q. From the given options: {answer_set}, select the correct answer (ONLY output the answer).  A. 2, 1, 0                      D. 2, 0, 1  B. 0, 2, 1                      E. 0, 1, 2  C. 1, 0, 2                      F. 1, 2, 0</p> <p><b>True/False</b></p> <p>Q. Given the predicted depth ordering as '2, 1, 0', evaluate the prediction as Correct or Incorrect.  <b>Ans. Correct</b></p>
	<p><b>Prompt</b></p> <p>The image shows 2D shapes placed randomly. The shapes overlap each other, creating a depth effect. When two shapes are overlapping, the shape that is complete is defined to be on top of the partially hidden shape. Each 2D shape has a number written over them which we call ShapeID and must be inferred as the label for the corresponding shape. Provide depth ordering from top to bottom for the shapes '3, 0, 1' in the image. Answer in the format: 'ShapeID, ShapeID, ...'. For eg. '3, 1, 2' is a valid answer format.</p>	<p><b>MCQ</b></p> <p>Q. From the given options: {answer_set}, select the correct answer (ONLY output the answer).  A. 1, 3, 0                      D. 3, 0, 1  B. 0, 3, 1                      E. 0, 1, 3  C. 1, 0, 3                      F. 3, 1, 0</p> <p><b>True/False</b></p> <p>Q. Given the predicted depth ordering as '3, 1, 0', evaluate the prediction as Correct or Incorrect.  <b>Ans. Correct</b></p>
	<p><b>Prompt</b></p> <p>The image shows 2D shapes placed randomly. The shapes overlap each other, creating a depth effect. When two shapes are overlapping, the shape that is complete is defined to be on top of the partially hidden shape. Each 2D shape has a number written over them which we call ShapeID and must be inferred as the label for the corresponding shape. Provide depth ordering from top to bottom for the shapes '26, 61' in the image. Answer in the format: 'ShapeID, ShapeID, ...'. For eg. '3, 1, 2' is a valid answer format.</p>	<p><b>MCQ</b></p> <p>Q. From the given options: {answer_set}, select the correct answer (ONLY output the answer).  A. 61, 26                      B. 26, 61</p> <p><b>True/False</b></p> <p>Q. Given the predicted depth ordering as '26, 61', evaluate the prediction as Correct or Incorrect.  <b>Ans. Incorrect</b></p>
	<p><b>Prompt</b></p> <p>The image shows 2D shapes placed randomly. The shapes overlap each other, creating a depth effect. When two shapes are overlapping, the shape that is complete is defined to be on top of the partially hidden shape. Each 2D shape has a unique color which we call the ShapeColor for the corresponding shape. Provide depth ordering from top to bottom for the shapes 'brown rectangle, blue triangle, pink triangle' in the image. Answer in the format: 'ShapeColor shape, ShapeColor shape, ...'. For eg. 'red triangle, blue circle, green rectangle' is a valid answer format.</p>	<p><b>MCQ</b></p> <p>Q. From the given options: {answer_set}, select the correct answer (ONLY output the answer).  A. brown rectangle, blue triangle, pink triangle  B. blue triangle, brown rectangle, pink triangle  C. brown rectangle, pink triangle, blue triangle  D. blue triangle, pink triangle, brown rectangle  E. pink triangle, blue triangle, brown rectangle  F. pink triangle, brown rectangle, blue triangle</p> <p><b>True/False</b></p> <p>Q. Given the predicted depth ordering as 'pink triangle, brown rectangle, blue triangle', evaluate the prediction as Correct or Incorrect.  <b>Ans. Incorrect</b></p>
	<p><b>Prompt</b></p> <p>The image shows 2D shapes placed randomly. The shapes overlap each other, creating a depth effect. When two shapes are overlapping, the shape that is complete is defined to be on top of the partially hidden shape. Each 2D shape has a unique color which we call the ShapeColor for the corresponding shape. Provide depth ordering from top to bottom for the shapes 'brown rectangle, cyan triangle' in the image. Answer in the format: 'ShapeColor shape, ShapeColor shape, ...'. For eg. 'red triangle, blue circle, green rectangle' is a valid answer format.</p>	<p><b>MCQ</b></p> <p>Q. From the given options: {answer_set}, select the correct answer (ONLY output the answer).  A. brown rectangle, cyan triangle  B. cyan triangle, brown rectangle</p> <p><b>True/False</b></p> <p>Q. Given the predicted depth ordering as '26, 61', evaluate the prediction as Correct or Incorrect.  <b>Ans. Incorrect</b></p>
	<p><b>Prompt</b></p> <p>The image shows 2D shapes placed randomly. The shapes overlap each other, creating a depth effect. When two shapes are overlapping, the shape that is complete is defined to be on top of the partially hidden shape. Each 2D shape has a unique color which we call the ShapeColor for the corresponding shape. Provide depth ordering from top to bottom for the shapes 'orange circle, brown triangle, magenta triangle' in the image. Answer in the format: 'ShapeColor shape, ShapeColor shape, ...'. For eg. 'red triangle, blue circle, green rectangle' is a valid answer format.</p>	<p><b>MCQ</b></p> <p>Q. From the given options: {answer_set}, select the correct answer (ONLY output the answer).  A. orange circle, brown triangle, magenta triangle  B. orange circle, magenta triangle, brown triangle  C. brown triangle, orange circle, magenta triangle  D. brown triangle, magenta triangle, orange circle  E. magenta triangle, orange circle, brown triangle  F. magenta triangle, brown triangle, orange circle</p> <p><b>True/False</b></p> <p>Q. Given the predicted depth ordering as 'brown triangle, magenta triangle, orange circle', evaluate the prediction as Correct or Incorrect.  <b>Ans. Incorrect</b></p>

Figure 11. **Samples from GeoMeter-2D dataset - depth category.** Here each row represents one image and its corresponding prompt along with MCQ and True/False questions. First three rows show samples for labels as query attribute, whereas last three rows show samples for color as query attribute.



**Prompt**

The image shows red 2D rectangles stacked on top of each other There are multiple stacks in the image. The black region at the bottom of the image is the ground level, and is where the base of the stack lies. The height of each stack is measured from its base. Each 2D shape has a number written over them which we call ShapeID and must be inferred as the label for the corresponding shape. The stacks are labelled A, B, C from left to right. Swap shape 1 from stack A with shape 9 from stack C. Now order the stacks labelled 'B, A, C' from shortest to tallest. Answer in the format: 'StackLabel, StackLabel, ...'. For eg. 'B, A, C' is a valid answer format.

**MCQ**

Q. From the given options: {answer\_set}, select the correct answer (ONLY output the answer).  
A. B, C, A D. A, C, B  
B. C, B, A E. B, A, C  
C. C, A, B F. A, B, C

**True/False**

Q. Given the predicted depth ordering as 'C, B, A', evaluate the prediction as Correct or Incorrect.  
Ans. Incorrect



**Prompt**

The image shows red 2D rectangles stacked on top of each other There are multiple stacks in the image. The black region at the bottom of the image is the ground level, and is where the base of the stack lies. The height of each stack is measured from its base. Each 2D shape has a number written over them which we call ShapeID and must be inferred as the label for the corresponding shape. The stacks are labelled A, B, C, D, E from left to right. Swap shape 2 from stack A with shape 15 from stack D. Now order the stacks labelled 'B, C, A' from shortest to tallest. Answer in the format: 'StackLabel, StackLabel, ...'. For eg. 'B, A, C' is a valid answer format.

**MCQ**

Q. From the given options: {answer\_set}, select the correct answer (ONLY output the answer).  
A. B, C, A D. A, C, B  
B. C, B, A E. B, A, C  
C. C, A, B F. A, B, C

**True/False**

Q. Given the predicted depth ordering as 'A, B, C', evaluate the prediction as Correct or Incorrect.  
Ans. Correct



**Prompt**

The image shows red 2D rectangles stacked on top of each other There are multiple stacks in the image. The black region at the bottom of the image is the ground level, and is where the base of the stack lies. The height of each stack is measured from its base. Each 2D shape has a number written over them which we call ShapeID and must be inferred as the label for the corresponding shape. The stacks are labelled A, B, C from left to right. Swap shape 4 from stack B with shape 9 from stack C. Now order the stacks labelled 'A, B, C' from shortest to tallest. Answer in the format: 'StackLabel, StackLabel, ...'. For eg. 'B, A, C' is a valid answer format.

**MCQ**

Q. From the given options: {answer\_set}, select the correct answer (ONLY output the answer).  
A. B, C, A D. A, C, B  
B. C, B, A E. B, A, C  
C. C, A, B F. A, B, C

**True/False**

Q. Given the predicted depth ordering as 'C, B, A', evaluate the prediction as Correct or Incorrect.  
Ans. Incorrect



**Prompt**

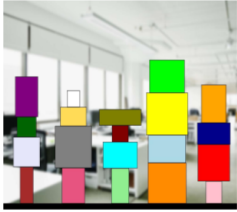
The image shows 2D rectangles stacked on top of each other There are multiple stacks in the image. The black region at the bottom of the image is the ground level, and is where the base of the stack lies. The height of each stack is measured from its base. Each 2D shape has a unique color. The stacks are labelled A, B, C from left to right. Swap light blue rectangle from stack B with light green rectangle from stack C. Now order the stacks labelled 'B, A, C' from shortest to tallest. Answer in the format: 'StackLabel, StackLabel, ...'. For eg. 'B, A, C' is a valid answer format.

**MCQ**

Q. From the given options: {answer\_set}, select the correct answer (ONLY output the answer).  
A. B, C, A D. A, C, B  
B. C, B, A E. B, A, C  
C. C, A, B F. A, B, C

**True/False**

Q. Given the predicted depth ordering as 'C, B, A', evaluate the prediction as Correct or Incorrect.  
Ans. Incorrect



**Prompt**

The image shows 2D rectangles stacked on top of each other There are multiple stacks in the image. The black region at the bottom of the image is the ground level, and is where the base of the stack lies. The height of each stack is measured from its base. Each 2D shape has a unique color. The stacks are labelled A, B, C, D, E from left to right. Swap red rectangle from stack E with navy blue rectangle from stack E. Now order the stacks labelled 'D, E, C' from shortest to tallest. Answer in the format: 'StackLabel, StackLabel, ...'. For eg. 'B, A, C' is a valid answer format.

**MCQ**

Q. From the given options: {answer\_set}, select the correct answer (ONLY output the answer).  
A. D, C, E D. E, C, D  
B. C, D, E E. D, E, C  
C. C, E, D F. E, D, C

**True/False**

Q. Given the predicted depth ordering as 'C, E, D', evaluate the prediction as Correct or Incorrect.  
Ans. Correct



**Prompt**

The image shows 2D rectangles stacked on top of each other There are multiple stacks in the image. The black region at the bottom of the image is the ground level, and is where the base of the stack lies. The height of each stack is measured from its base. Each 2D shape has a unique color. The stacks are labelled A, B, C from left to right. Swap dark green rectangle from stack C with orange rectangle from stack C. Now order the stacks labelled 'C, A, B' from shortest to tallest. Answer in the format: 'StackLabel, StackLabel, ...'. For eg. 'B, A, C' is a valid answer format.

**MCQ**

Q. From the given options: {answer\_set}, select the correct answer (ONLY output the answer).  
A. B, C, A D. C, A, B  
B. B, A, C E. C, B, A  
C. A, C, B F. A, B, C

**True/False**

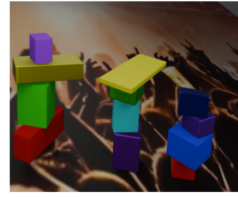
Q. Given the predicted depth ordering as 'A, B, C', evaluate the prediction as Correct or Incorrect.  
Ans. Correct

Figure 12. Samples from GeoMeter-2D dataset - height category. Here each row represents one image and its corresponding prompt along with MCQ and True/False questions. First three rows show samples for labels as query attribute, whereas last three rows show samples for color as query attribute



	<p><b>Prompt</b></p> <p>The image shows 3D shapes placed randomly. From the camera viewpoint distance some shapes are in front and some are at the back, creating a depth effect. Each 3D shape has a color and an associated material which we call Color and Material and must be inferred as the label for the corresponding shape. Provide depth ordering from front to back for the shapes <b>'red cube, purple cylinder'</b> in the image. Answer in the format: 'ShapeColor shape, ShapeColor shape, ...'. For eg. 'red cube, blue sphere, green cylinder is a valid answer format.</p>	<p><b>MCQ</b></p> <p>Q. From the given options: (<b>answer_set</b>), select the correct answer (ONLY output the answer).  <b>A. red cube, purple cylinder</b>  <b>B. purple cylinder, red cube</b></p> <p><b>True/False</b></p> <p>Q. Given the predicted depth ordering as <b>'purple cylinder, red cube'</b>, evaluate the prediction as Correct or Incorrect.  <b>Ans. Incorrect</b></p>
	<p><b>Prompt</b></p> <p>The image shows 3D shapes placed randomly. From the camera viewpoint distance some shapes are in front and some are at the back, creating a depth effect. Each 3D shape has a color and an associated material which we call Color and Material and must be inferred as the label for the corresponding shape. Provide depth ordering from front to back for the shapes <b>'red cylinder, green cube, yellow cylinder'</b> in the image. Answer in the format: 'ShapeColor shape, ShapeColor shape, ...'. For eg. 'red cube, blue sphere, green cylinder is a valid answer format.</p>	<p><b>MCQ</b></p> <p>Q. From the given options: (<b>answer_set</b>), select the correct answer (ONLY output the answer).  <b>A. red cylinder, green cube, yellow cylinder</b>  <b>B. red cylinder, yellow cylinder, green cube</b>  <b>C. green cube, red cylinder, yellow cylinder</b>  <b>D. green cube, yellow cylinder, red cylinder</b>  <b>E. yellow cylinder, red cylinder, green cube</b>  <b>F. yellow cylinder, green cube, red cylinder</b></p> <p><b>True/False</b></p> <p>Q. Given the predicted depth ordering as <b>'yellow cylinder, red cylinder, green cube'</b>, evaluate the prediction as Correct or Incorrect.  <b>Ans. Incorrect</b></p>
	<p><b>Prompt</b></p> <p>The image shows 3D shapes placed randomly. From the camera viewpoint distance some shapes are in front and some are at the back, creating a depth effect. Each 3D shape has a color and an associated material which we call Color and Material and must be inferred as the label for the corresponding shape. Provide depth ordering from front to back for the shapes <b>'red cylinder, purple sphere, green cylinder'</b> in the image. Answer in the format: 'ShapeColor shape, ShapeColor shape, ...'. For eg. 'red cube, blue sphere, green cylinder is a valid answer format.</p>	<p><b>MCQ</b></p> <p>Q. From the given options: (<b>answer_set</b>), select the correct answer (ONLY output the answer).  <b>A. red cylinder, purple sphere, green cylinder</b>  <b>B. red cylinder, green cylinder, purple sphere</b>  <b>C. purple sphere, red cylinder, green cylinder</b>  <b>D. purple sphere, green cylinder, red cylinder</b>  <b>E. green cylinder, red cylinder, purple sphere</b>  <b>F. green cylinder, purple sphere, red cylinder</b></p> <p><b>True/False</b></p> <p>Q. Given the predicted depth ordering as <b>green cylinder, red cylinder, purple sphere</b>, evaluate the prediction as Correct or Incorrect.  <b>Ans. Correct</b></p>
	<p><b>Prompt</b></p> <p>The image shows 3D shapes placed randomly. From the camera viewpoint distance some shapes are in front and some are at the back, creating a depth effect. Each 3D shape has a color and an associated material which we call Color and Material and must be inferred as the label for the corresponding shape. Provide depth ordering from front to back for the shapes <b>'red rubber cube, cyan rubber sphere'</b> in the image. Answer in the format: 'ShapeColor ShapeMaterial shape, ShapeColor ShapeMaterial shape, ...'. For eg. 'red metal cube, blue rubber sphere, green metal cylinder is a valid answer format.</p>	<p><b>MCQ</b></p> <p>Q. From the given options: (<b>answer_set</b>), select the correct answer (ONLY output the answer).  <b>A. red rubber cube, cyan rubber cylinder</b>  <b>B. cyan rubber cylinder, red rubber cube</b></p> <p><b>True/False</b></p> <p>Q. Given the predicted depth ordering as <b>red rubber cube, cyan rubber cylinder</b>, evaluate the prediction as Correct or Incorrect.  <b>Ans. Correct</b></p>
	<p><b>Prompt</b></p> <p>The image shows 3D shapes placed randomly. From the camera viewpoint distance some shapes are in front and some are at the back, creating a depth effect. Each 3D shape has a color and an associated material which we call Color and Material and must be inferred as the label for the corresponding shape. Provide depth ordering from front to back for the shapes <b>'red metal sphere, blue rubber cube'</b> in the image. Answer in the format: 'ShapeColor ShapeMaterial shape, ShapeColor ShapeMaterial shape, ...'. For eg. 'red metal cube, blue rubber sphere, green metal cylinder is a valid answer format.</p>	<p><b>MCQ</b></p> <p>Q. From the given options: (<b>answer_set</b>), select the correct answer (ONLY output the answer).  <b>A. red metal sphere, blue rubber cube</b>  <b>B. blue rubber cube, red metal sphere</b></p> <p><b>True/False</b></p> <p>Q. Given the predicted depth ordering as <b>blue rubber cube, red metal sphere</b> evaluate the prediction as Correct or Incorrect.  <b>Ans. Correct</b></p>
	<p><b>Prompt</b></p> <p>The image shows 3D shapes placed randomly. From the camera viewpoint distance some shapes are in front and some are at the back, creating a depth effect. Each 3D shape has a color and an associated material which we call Color and Material and must be inferred as the label for the corresponding shape. Provide depth ordering from front to back for the shapes <b>'green rubber sphere, purple metal sphere, blue rubber cylinder'</b> in the image. Answer in the format: 'ShapeColor ShapeMaterial shape, ShapeColor ShapeMaterial shape, ...'. For eg. 'red metal cube, blue rubber sphere, green metal cylinder is a valid answer format.</p>	<p><b>MCQ</b></p> <p>Q. From the given options: (<b>answer_set</b>), select the correct answer (ONLY output the answer).  <b>A. green rubber sphere, purple metal sphere, blue rubber cylinder</b>  <b>B. green rubber sphere, blue rubber cylinder, purple metal sphere</b>  <b>C. purple metal sphere, green rubber sphere, blue rubber cylinder</b>  <b>D. purple metal sphere, blue rubber cylinder, green rubber sphere</b>  <b>E. blue rubber cylinder, green rubber sphere, purple metal sphere</b>  <b>F. blue rubber cylinder, purple metal sphere, green rubber sphere</b></p> <p><b>True/False</b></p> <p>Q. Given the predicted depth ordering as <b>purple metal sphere, green rubber sphere, blue rubber cylinder</b> evaluate the prediction as Correct or Incorrect.  <b>Ans. Incorrect</b></p>

Figure 13. Samples from GeoMeter-3D dataset - depth category. Here each row represents one image and its corresponding prompt along with MCQ and True/False questions. First three rows show samples for color as query attribute, whereas last three rows show samples for color+material as query attribute



**Prompt**

The image shows 3D cubes stacked on top of each other. There are multiple stacks in the image. If there is a black cube at the bottom of the stack, then that is considered as the ground level, and the stack lies on top of it. The height of each stack is measured from its base. Each 3D shape has a unique color. The stacks are labelled A, B, C,... from left to right. Swap cyan cube from stack B with red cube from stack A. Now order the stacks labelled 'B, A, C' from shortest to tallest. Answer in the format: 'StackLabel, StackLabel, ...'. For eg. 'B, A, C' is a valid answer format.

**MCQ**

Q. From the given options: {answer\_set}, select the correct answer (ONLY output the answer).

A. B, C, A      D. A, C, B  
B. C, B, A      E. B, A, C  
C. C, A, B      F. A, B, C

**True/False**

Q. Given the predicted depth ordering as 'C, B, A', evaluate the prediction as Correct or Incorrect.

Ans. Incorrect



**Prompt**

The image shows 3D cubes stacked on top of each other. There are multiple stacks in the image. If there is a black cube at the bottom of the stack, then that is considered as the ground level, and the stack lies on top of it. The height of each stack is measured from its base. Each 3D shape has a unique color. The stacks are labelled A, B, C,... from left to right. Swap purple cube from stack A with blue cube from stack C. Now order the stacks labelled 'A, C, B' from shortest to tallest. Answer in the format: 'StackLabel, StackLabel, ...'. For eg. 'B, A, C' is a valid answer format.

**MCQ**

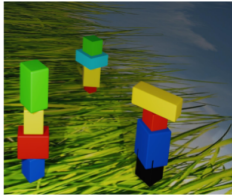
Q. From the given options: {answer\_set}, select the correct answer (ONLY output the answer).

A. B, C, A      D. A, C, B  
B. C, B, A      E. B, A, C  
C. C, A, B      F. A, B, C

**True/False**

Q. Given the predicted depth ordering as 'C, B, A', evaluate the prediction as Correct or Incorrect.

Ans. Incorrect



**Prompt**

The image shows 3D cubes stacked on top of each other. There are multiple stacks in the image. If there is a black cube at the bottom of the stack, then that is considered as the ground level, and the stack lies on top of it. The height of each stack is measured from its base. Each 3D shape has a unique color. The stacks are labelled A, B, C,... from left to right. Swap red cube from stack A with cyan cube from stack B. Now order the stacks labelled 'A, B, C' from shortest to tallest. Answer in the format: 'StackLabel, StackLabel, ...'. For eg. 'B, A, C' is a valid answer format.

**MCQ**

Q. From the given options: {answer\_set}, select the correct answer (ONLY output the answer).

A. B, C, A      D. A, C, B  
B. C, B, A      E. B, A, C  
C. C, A, B      F. A, B, C

**True/False**

Q. Given the predicted depth ordering as 'C, B, A', evaluate the prediction as Correct or Incorrect.

Ans. Incorrect



**Prompt**

The image shows 3D cubes stacked on top of each other. There are multiple stacks in the image. If there is a black cube at the bottom of the stack, then that is considered as the ground level, and the stack lies on top of it. The height of each stack is measured from its base. Each 3D shape has a unique color and material. The stacks are labelled A, B, C,... from left to right. Swap cyan rubber cube from stack A with cyan metal cube from stack E. Now order the stacks labelled 'A, B, C' from shortest to tallest. Answer in the format: 'StackLabel, StackLabel, ...'. For eg. 'B, A, C' is a valid answer format.

**MCQ**

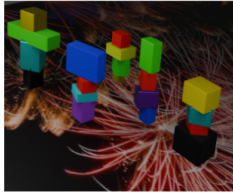
Q. From the given options: {answer\_set}, select the correct answer (ONLY output the answer).

A. B, C, A      D. A, C, B  
B. C, B, A      E. B, A, C  
C. C, A, B      F. A, B, C

**True/False**

Q. Given the predicted depth ordering as 'C, B, A', evaluate the prediction as Correct or Incorrect.

Ans. Incorrect



**Prompt**

The image shows 3D cubes stacked on top of each other. There are multiple stacks in the image. If there is a black cube at the bottom of the stack, then that is considered as the ground level, and the stack lies on top of it. The height of each stack is measured from its base. Each 3D shape has a unique color and material. The stacks are labelled A, B, C,... from left to right. Swap green rubber cube from stack A with green rubber cube from stack D. Now order the stacks labelled 'A, D, E' from shortest to tallest. Answer in the format: 'StackLabel, StackLabel, ...'. For eg. 'B, A, C' is a valid answer format.

**MCQ**

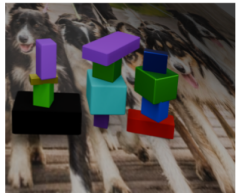
Q. From the given options: {answer\_set}, select the correct answer (ONLY output the answer).

A. A, D, E      D. D, E, A  
B. A, E, D      E. E, A, D  
C. D, A, E      F. E, D, A

**True/False**

Q. Given the predicted depth ordering as 'D, E, A', evaluate the prediction as Correct or Incorrect.

Ans. Incorrect



**Prompt**

The image shows 3D cubes stacked on top of each other. There are multiple stacks in the image. If there is a black cube at the bottom of the stack, then that is considered as the ground level, and the stack lies on top of it. The height of each stack is measured from its base. Each 3D shape has a unique color and material. The stacks are labelled A, B, C,... from left to right. Swap green rubber cube from stack A with green rubber cube from stack B. Now order the stacks labelled 'A, C, B' from shortest to tallest. Answer in the format: 'StackLabel, StackLabel, ...'. For eg. 'B, A, C' is a valid answer format.

**MCQ**

Q. From the given options: {answer\_set}, select the correct answer (ONLY output the answer).

A. B, C, A      D. A, C, B  
B. C, B, A      E. B, A, C  
C. C, A, B      F. A, B, C

**True/False**

Q. Given the predicted depth ordering as 'A, C, B', evaluate the prediction as Correct or Incorrect.

Ans. Correct

Figure 14. Samples from GeoMeter-3D dataset - height category. Here each row represents one image and its corresponding prompt along with MCQ and True/False questions. First three rows show samples for color as query attribute, whereas last three rows show samples for color+material as query attribute

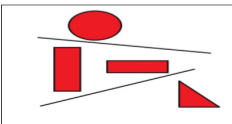
	<p><b>Prompt</b></p> <p>How many pairs of lines in the image are parallel?</p>	<p><b>MCQ</b></p> <p>Q. From the given options: {answer_set}, select the correct answer (ONLY output the answer).  A. 2  B. 1  C. 3  D. 4</p>
	<p><b>Prompt</b></p> <p>Is the black line perpendicular to the blue line?</p>	<p><b>MCQ</b></p> <p>Q. From the given options: {answer_set}, select the correct answer (ONLY output the answer).  A. (Yes)                      B. (No)</p>
<hr/>		
	<p><b>Prompt</b></p> <p>Which are the different kinds of shapes seen in the image? Answer in the format 'shape1, shape2...'. For example, 'triangle, circle' is a valid answer format.</p>	<p><b>MCQ</b></p> <p>Q. From the given options: {answer_set}, select the correct answer (ONLY output the answer).  A. circle, rectangle  B. rectangle, triangle  C. circle, rectangle, triangle  D. circle, square</p>
	<p><b>Prompt</b></p> <p>Which are the different kinds of shapes seen in the image? Answer in the format 'shape1, shape2...'. For example, 'triangle, circle' is a valid answer format.</p>	<p><b>MCQ</b></p> <p>Q. From the given options: {answer_set}, select the correct answer (ONLY output the answer).  A. circle, rectangle, triangle  B. circle  C. circle, square  D. square, triangle</p>
<hr/>		
	<p><b>Prompt</b></p> <p>The image shows various shapes. How many triangles are in the image?</p>	<p><b>MCQ</b></p> <p>Q. From the given options: {answer_set}, select the correct answer (ONLY output the answer).  A. 2  B. 1  C. 3  D. 4</p>
	<p><b>Prompt</b></p> <p>The image shows various shapes. How many rectangles are in the image?</p>	<p><b>MCQ</b></p> <p>Q. From the given options: {answer_set}, select the correct answer (ONLY output the answer).  A. 2  B. 4  C. 3  D. 1</p>
<hr/>		
	<p><b>Prompt</b></p> <p>How many shapes are to the left of the red circle?</p>	<p><b>MCQ</b></p> <p>Q. From the given options: {answer_set}, select the correct answer (ONLY output the answer).  A. 2  B. 1  C. 3  D. 4</p>
	<p><b>Prompt</b></p> <p>The image shows some shapes and two lines. How many shapes are in between the two lines?</p>	<p><b>MCQ</b></p> <p>Q. From the given options: {answer_set}, select the correct answer (ONLY output the answer).  A. 2  B. 3  C. 1  D. 4</p>

Figure 15. **Samples from GeoMeter-2D-Basic dataset.** Here each two rows respectively represent line understanding, shape identification, shape counting and spatial relationship categories. Each row shows one image and its corresponding prompt along with the MCQ.