A. Appendix Outline

In the supplementary materials, we provide:

- A detailed description of the VisNumBench construction process (Appendix B);
- The evaluation setup and comprehensive results for VisNumBench sub-experiments (Appendix C);
- Additional visualizations (Appendix D).

B. Details of VisNumBench Construction

Angle The task may involve recognizing angles in both 2D and 3D contexts, such as the angles between intersecting lines or the angle between the viewpoint and an object. The figures for VisNumBench-Synthetic are either generated using Python programs or sourced from VisOnlyQA [20], whereas the images for VisNumBench-Real are either captured by the authors or collected from Google Images [17].

Length Based on both synthetic and real-world scenes, we designed a variety of question-answering tasks, including relative length comparison, multiple segment estimation, and the estimation of object length, height, and proportion, among others. The figures in VisNumBench-Synthetic are either generated using Python programs or sourced from MathVista [30], whereas the figures in VisNumBench-Real are captured by the authors or collected from Google Images [17].

Scale We provide figures illustrating the coordinates of a point in a coordinate system, the time indicated on a clock, and the temperature displayed on a thermometer. The figures for VisNumBench-Synthetic are generated by Python programs, sourced from MathVista [20], while other figures originate from Google Images [17] and ECharts [12]. In contrast, the statistics for VisNumBench-Real are either captured by the authors or collected from Google Images [17].

Quantity Each figure contains a varying number of objects, such as points or triangles in synthetic scenarios, or hot air balloons or pets in real-world scenarios. The figures for VisNumBench-Synthetic are either generated using Python programs or obtained from MathVista [30], whereas the figures for VisNumBench-Real are sourced from Google Images [17] and the ShanghaiTech dataset [57].

Depth We further refine the "Relative Depth" task in BLINK [16] by incorporating additional choice points and introducing new question-answer formats. MLLMs will be presented with images containing objects at varying depths, requiring them to determine the correct depth order or estimate the relative distances between objects.

The figures for VisNumBench-Synthetic are obtained from the WallpapersCraft website [43] or sourced from MathVista [30] and VSLAM-TartanAir [48]. The figures for VisNumBench-Real are sourced from BLINK [16] and the NYU Depth Dataset V2 [41].

Area VisNumBench-Synthetic contains comparisons and estimations of object area sizes for both identical and different shapes, as well as their multiplicative relationships. The figures in VisNumBench-Synthetic are either generated by Python programs or obtained from VisOnlyQA [20].

Volume Objects are presented from different perspectives and in various sizes, requiring MLLMs to infer relative volume sizes and proportions based on visible dimensions and depth. The figures for VisNumBench-Real are obtained through camera capture or sourced from Google Images [17].

More details of data construction are shown in the Tables 5, 6. Figure 8 shows the dataset statistics of VisNumBench based on various visual numerical estimation tasks. Figure 9 shows how to build a QA pair based on images generated by a Python script. Python-generated images are 500 × 500, and all others are resized with the longer side capped at 500 pixels.

Table 5. The source distribution of different visual numerical attributes on the VisNumBench-Synthetic set.

	Python Program	Web Collection	Other Dataset	Total
Angle	138	0	32	170
Length	160	0	21	181
Scale	77	50	13	140
Quantity	185	0	11	196
Depth	0	70	65	135
Area	139	0	50	189
Total	699	120	192	1011

Table 6. The source distribution of different visual numerical attributes on the VisNumBench-Real set.

	Image Taken by Us	Web Collection	Other Dataset	Total
Angle	91	58	0	149
Length	144	18	0	162
Scale	21	122	0	143
Quantity	0	113	34	147
Depth	0	0	154	154
Volume	140	7	0	147
Total	396	318	188	902

C. Evalution Details

C.1. Model Access

This section provides details on model access and parameter settings (refer to Table 7). The model responses presented in this paper were collected between January 1 and February 28, 2025. We set $max\_new\_tokens \geq 512$, while all other parameters were kept at their default values.

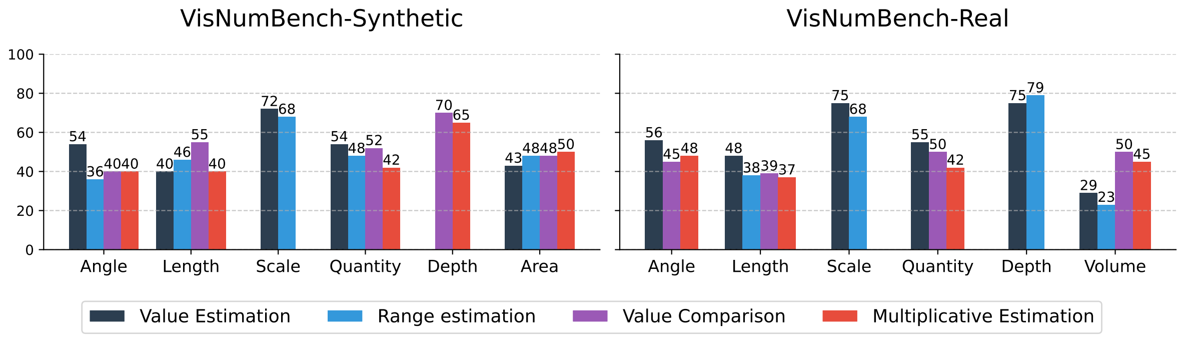
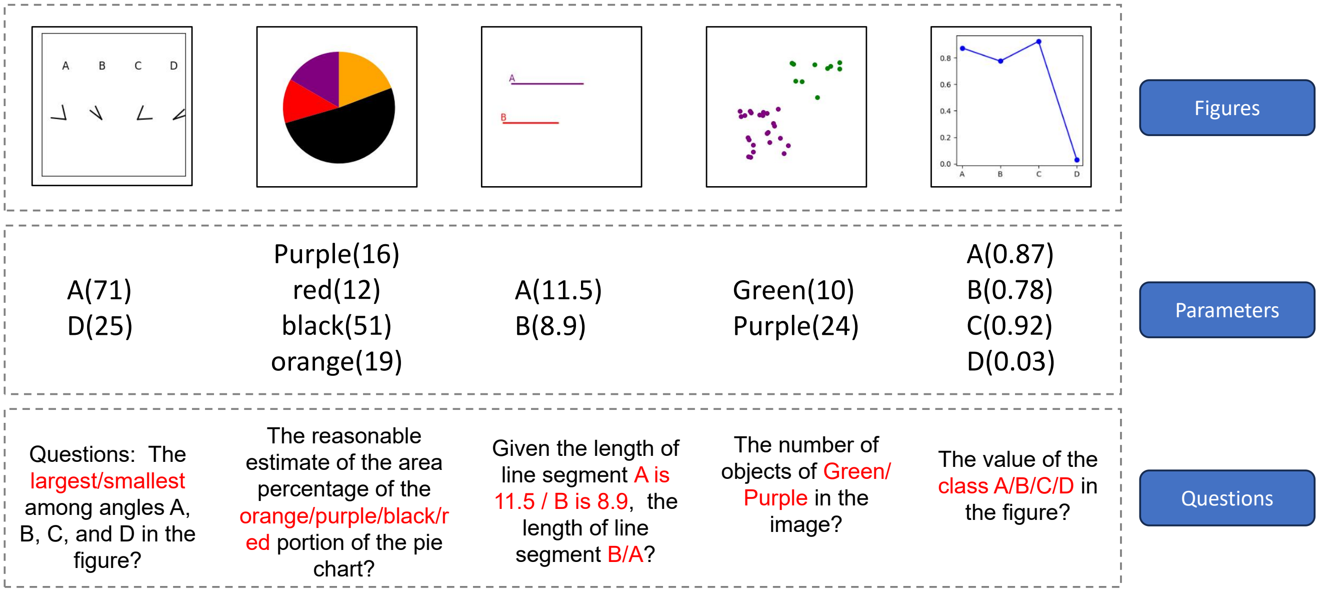


Figure 8. Dataset statistics of VisNumBench based on various visual numerical estimation tasks.



Questions: The largest/smallest among angles A, B, C, and D in the figure?

The reasonable estimate of the area percentage of the orange/purple/black/red portion of the pie chart?

Given the length of line segment A is 11.5 / B is 8.9, the length of line segment B/A?

The number of objects of Green/Purple in the image?

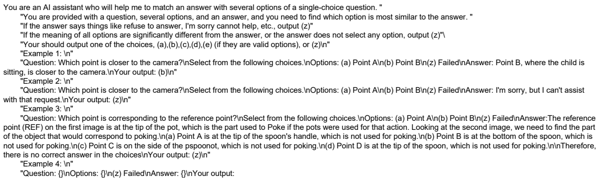
The value of the class A/B/C/D in the figure?

Figure 9. Examples of the data generated by Python and manually designed questions.

C.2. Detail of Evaluation

Prompt. Table 8 shows the prompts for evaluation. The one below is the prompt with CoT.

Post-processing. We use InternVL2.5-38B to extract the selected options from the MLLMs response, and the corresponding extracting prompt is referenced from [16], as shown in Figure 10.



```
You are an AI assistant who will help me to match an answer with several options of a single-choice question. "
        "You are provided with a question, several options, and an answer, and you need to find which option is most similar to the answer. "
        "If the answer says things like refuse to answer, I'm sorry cannot help, etc., output (z)"
        "If the meaning of all options are significantly different from the answer, or the answer does not select any option, output (z)"\
        "Your should output one of the choices, (a),(b),(c),(d),(e) (if they are valid options), or (z)\n"
        "Example 1: \n"
        "Question: Which point is closer to the camera?\nSelect from the following choices.\nOptions: (a) Point A\n(b) Point B\n(z) Failed\nAnswer: Point B, where the child is
sitting, is closer to the camera.\nYour output: (b)\n"
        "Example 2: \n"
        "Question: Which point is closer to the camera?\nSelect from the following choices.\nOptions: (a) Point A\n(b) Point B\n(z) Failed\nAnswer: I'm sorry, but I can't assist
with that request.\nYour output: (z)\n"
        "Example 3: \n"
        "Question: Which point is corresponding to the reference point?\nSelect from the following choices.\nOptions: (a) Point A\n(b) Point B\n(z) Failed\nAnswer:The reference
point (REF) on the first image is at the tip of the pot, which is the part used to Poke if the pots were used for that action. Looking at the second image, we need to find the part
of the object that would correspond to poking.\n(a) Point A is at the tip of the spoon's handle, which is not used for poking.\n(b) Point B is at the bottom of the spoon, which is
not used for poking.\n(c) Point C is on the side of the pspoonot, which is not used for poking.\n(d) Point D is at the tip of the spoon, which is not used for poking.\n\nTherefore,
there is no correct answer in the choices\nYour output: (z)\n"
        "Example 4: \n"
        "Question: {}\nOptions: {}\n(z) Failed\nAnswer: {}\nYour output:
```

Figure 10. Prompt for extracting selected options from the responses of MLLMs.

Human Evaluation. Two individuals with backgrounds in computer science, who were not involved in the project, independently participated by responding to the questions through a visualization interface. No compensation was provided for their participation.

C.3. Additional Results

This section provides additional results of experiments in Section 4.2.

Table 9 presents the improvements introduced by the CoT prompt. It can be observed that, except for Gemini 2.0 Flash, which shows a positive gain on VisNumBench-Synthetic, the accuracy of the other models decreases. Tables 10 and 11 report the results of multimodal mathematical models and multimodal CoT models, respectively, corresponding to Figure 7. Table 12 presents the performance of models of varying sizes from the QwenVL and InternVL

Table 7. The MLLMs evaluated in this paper. This table presents the model names (Hugging Face repository name or Official API name).

Phi-3.5-vision	microsoft/Phi-3.5-vision-instruct
LLaVA-v1.5-7B	liuhaotian/llava-v1.5-7b
LLaVA-v1.5-13B	liuhaotian/llava-v1.5-13b
LLaVA-v1.6-34B	liuhaotian/llava-v1.6-34b
LLaVA-Onevision-7B	llava-hf/llava-onevision-qwen2-72b-si-hf
LLaVA-Onevision-72B	llava-hf/llava-onevision-qwen2-72b-ov-hf
InternVL2.5-8B	OpenGVLab/InternVL2_5-8B
InternVL2.5-38B	OpenGVLab/InternVL2_5-38B
InternVL2.5-78B	OpenGVLab/InternVL2_5-78B
Janus-Pro-7B	deepseek-ai/Janus-Pro-7B
Qwen2.5-VL-3B	Qwen/Qwen2.5-VL-3B-Instruct
Qwen2.5-VL-7B	Qwen/Qwen2.5-VL-7B-Instruct
Qwen2.5-VL-72B	Qwen/Qwen2.5-VL-72B-Instruct
GPT-4o	gpt-4o-2024-08-06
Gemini 1.5 Flash	gemini-1.5-flash
Gemini 2.0 Flash	gemini-2.0-flash
Gemini 1.5 Pro	gemini-1.5-pro-002

Table 8. The prompt employed for the evaluation of the benchmark.

Prompt

Question: {QUESTION}
Options: {OPTIONS}
Answer the question based on the most likely options.
Provide only the letter corresponding to your choice as the answer (e.g., '(a)', '(b)', '(c)', '(d)', '(e)').

Question: {QUESTION}
Options: {OPTIONS}
Please think and answer the question based on the most likely options.

families, extending the results shown in Table 4. Table 13 summarizes the overall accuracy of various models on the VisNumBench benchmark, covering both synthetic and real subsets.

C.4. Error Analysis

We use the results of Gemini 2.0 Flash as a case study for error analysis. Based on these results, we randomly selected 10 erroneous instances for each attribute in each scenario, manually analyzing a total of 120 randomly sampled errors across all tasks. We categorize the errors into two types: (1) Errors arising from the model's failure to accurately perceive the image (*image perception errors*); (2) Errors in which the model correctly interprets both the image and the question but fails to produce the correct numerical answer (*numerical intuition errors*). Our analysis reveals that 28.3% of the errors are due to image perception issues, particularly in scale-related tasks, where the model struggles to identify the positions of pointers. The remaining 71.7% are attributed to numerical intuition errors, which commonly involve difficulties with depth estimation, angle relationships, and quantity perception. These findings further substantiate that current models indeed lack robust numerical intuition.

D. Example Data and Model Outputs

Figures 11 to 22 show examples from VisNumBench and the responses of Gemini 2.0 Flash.

Table 9. The experiment results of the state-of-the-art MLLMs. with CoT prompt.

Models	Angle	Length	Scale	Quantity	Depth	Area/Volume	Average
VisNumBench-Systentic							
InternVL2.5-78B	27.06	51.93	68.57	46.43	51.85	74.60	53.21
Qwen2.5-VL-72B	38.24	52.49	68.57	56.63	48.89	74.07	56.68
Gemini 2.0 Flash	34.12	55.80	85.00	58.16	55.56	74.60	60.14
VisNumBench-Real							
InternVL2.5-78B	30.87	59.26	58.74	78.23	48.70	55.10	55.10
Qwen2.5-VL-72B	32.21	52.47	52.45	70.75	46.10	53.06	51.11
Gemini 2.0 Flash	33.56	49.38	72.03	80.95	41.56	59.86	55.88

Table 10. Accuracies of multimodal mathematical models, multimodal CoT models, and their respective base models (before fine-tuning) on VisNumBench-Synthetic.

Models	Angle	Length	Scale	Quantity	Depth	Area	Average
multimodal mathematical models							
Internvl-8B	28.24	49.72	55.00	28.57	31.11	46.03	39.56
InternVL2-8B-MPO	22.94	48.07	63.57	33.67	28.15	48.68	40.65
LLaVA-v1.5-13B	35.88	30.94	32.14	36.73	33.33	24.34	32.15
Math-LLaVA-13B	31.76	45.30	28.57	39.80	37.78	30.16	35.81
multimodal CoT models							
Llama-VL-3_2-11B	29.41	41.44	58.57	47.96	42.96	44.97	43.92
Llama-3.2V-11B-cot	28.82	46.41	40.71	56.12	38.52	57.14	45.50
Qwen2.5-VL-7B-Instruct	23.53	53.59	55.00	39.29	48.89	58.20	46.19
R1-Onevision-7B	33.53	37.57	56.43	34.69	23.70	47.09	38.87

Table 11. Accuracies of multimodal mathematical models, multimodal CoT models, and their respective base models (before fine-tuning) on VisNumBench-Real.

	Angle	Length	Scale	Quantity	Depth	Area	Average
multimodal mathematical models							
Internvl-8B	30.87	36.42	29.37	71.43	30.52	39.46	39.58
InternVL2-8B-MPO	30.87	35.19	29.37	72.11	34.42	38.10	39.91
LLaVA-v1.5-13B	28.86	43.21	29.37	46.94	49.35	41.50	40.02
Math-LLaVA-13B	20.13	35.80	23.78	45.58	37.66	35.37	33.15
multimodal CoT models							
Llama-VL-3_2-11B	38.26	40.74	30.77	69.39	38.31	42.18	43.24
Llama-3.2V-11B-cot	27.52	38.89	23.08	64.63	31.82	44.22	38.36
Qwen2.5-VL-7B-Instruct	24.16	38.89	32.17	59.18	48.70	42.86	41.02
R1-Onevision-7B	28.86	39.51	44.76	51.70	32.47	32.65	38.25

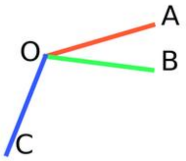
Table 12. Results of models with varying sizes from the QwenVL family and InternVL family on VisNumBench.

	Angle	Length	Scale	Quantity	Depth	Area/Volume	Average
VisNumBench-Systentic							
Qwen2-VL-2B	28.24	30.39	35.00	35.20	21.48	38.10	31.85
Qwen2-VL-7B	27.06	45.30	55.00	44.39	34.81	46.56	42.24
Qwen2-VL-72B	32.94	57.46	63.57	54.59	58.52	59.79	54.20
Internvl-8B	28.24	49.72	55.00	28.57	31.11	46.03	39.56
Internvl-40B	23.53	58.56	57.14	37.24	37.78	58.20	45.50
Qwen2.5-VL-3B	30.00	49.17	50.71	32.14	42.22	51.85	42.43
Qwen2.5-VL-7B	23.53	53.59	55.00	39.29	48.89	58.20	46.19
Qwen2.5-VL-72B	37.06	59.67	65.00	57.65	61.48	70.37	58.46
InternVL2.5-8B	26.47	41.99	49.29	34.69	41.48	46.03	39.66
InternVL2.5-38B	39.41	59.67	59.29	54.08	60.74	61.38	55.59
VisNumBench-Real							
Qwen2-VL-2B	10.74	19.75	19.58	47.62	32.47	19.73	24.94
Qwen2-VL-7B	19.46	38.89	30.07	67.35	41.56	54.42	41.91
Qwen2-VL-72B	21.48	45.06	37.06	74.83	48.70	52.38	46.56
Internvl-8B	30.87	36.42	29.37	71.43	30.52	39.46	39.58
Internvl-40B	30.87	50.00	28.67	72.79	35.71	52.38	45.12
Qwen2.5-VL-3B	30.20	44.44	35.66	51.70	43.51	49.66	42.57
Qwen2.5-VL-7B	24.16	38.89	32.17	59.18	48.70	42.86	41.02
Qwen2.5-VL-72B	34.23	50.62	43.36	80.27	52.60	59.18	53.33
InternVL2.5-8B	28.86	34.57	15.38	64.63	49.35	47.62	40.13
InternVL2.5-38B	30.20	51.85	26.57	83.67	61.04	58.50	52.11

Table 13. Performance (%) of various models on VisNumBench-Synthetic, VisNumBench-Real, and the overall VisNumBench.

Model	**Synthetic (1,011)**	**Real (902)**	**Overall**
Human	95.33	97.33	96.27
Gemini 2.0 Flash	57.57	**56.54**	**57.08**
InternVL2.5-78B	56.18	**56.54**	56.35
Qwen2.5-VL-72B	**58.46**	53.33	56.04
InternVL2.5-38B	55.59	52.11	53.95
LLaVA-Onevision-72B	50.84	50.78	50.81
Qwen2-VL-72B	54.20	46.56	50.60
LLaVA-v1.6-34B	44.31	50.55	47.25
Gemini 1.5 Pro	44.02	48.67	46.21
InternVL2-40B	45.50	45.12	45.32
Qwen2.5-VL-7B	46.19	41.02	43.75
Llama-VL-3_2-11B	43.92	43.24	43.60
Qwen2.5-VL-3B	42.43	42.57	42.50
Qwen2-VL-7B	42.24	41.91	42.08
Llama-3.2V-11B-cot	45.50	38.36	42.13
GPT-4o	43.72	39.58	41.77
InternVL2-8B-MPO	40.65	39.91	40.30
LLaVA-Onevision-7B	39.96	40.58	40.25
InternVL2.5-8B	39.66	40.13	39.88
InternVL2-8B	39.56	39.58	39.57
R1-Onevision-7B	38.87	38.25	38.58
LLaVA-v1.5-13B	32.15	40.02	35.86
Janus-Pro-7B	37.69	34.26	36.07
Phi-3.5-vision	32.34	37.25	34.66
Math-LLaVA-13B	35.81	33.15	34.56
Gemini 1.5 Flash	33.33	33.70	33.50
LLaVA-v1.5-7B	29.38	28.49	28.96
Qwen2-VL-2B	31.85	24.94	28.59
Random	24.76	25.54	25.13

Synthetic (Angle)



Prompt:
Question: Angle AOC is X times larger than Angel AOB. Which of the following options is a reasonable estimate for X?
Options: (a) 1.5 (b) 5.6 (c) 3.1 (d) 4.0
Please think and answer the question based on the most likely options.

Type: Multiplicative Estimation

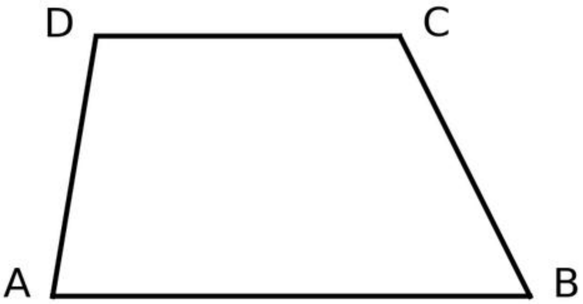
Answer:(b) 5.6

Gemini 2.0 Flash:

Based on the image, angle AOC appears to be significantly larger than angle AOB. If angle AOC were about 3 times larger than angle AOB, then 3.1 (c) would be a reasonable estimate for X. 4.0 (d) would also be a reasonable estimate. Therefore, considering the options, and the size of AOB and AOC, the most reasonable estimate for x would be 4.0. So the answer is (d).

⊗

Synthetic (Angle)



Prompt:
Question: Which angle is the largest among angles A, B, C, and D in the figure?
Options:(a) angle A (b) angle B (c) angle C (d) angle D
Please think and answer the question based on the most likely options.

Type: Value Comparison

Answer: (c) angle C

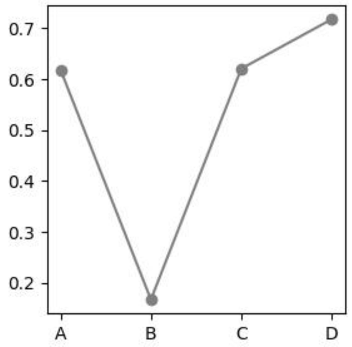
Gemini 2.0 Flash:

Based on the figure, angle B appears to be the largest angle. So the answer is (b).

⊗

Figure 11. Examples of VisNumBench and the results predicted by Gemini2.0 Flash (VisNumBench-Synthetic, Angle, 1/12).

Synthetic (Scale)



Prompt:
Question: Which of the following ranges is most likely to be the value of the class A in the figure?
Options: (a) (0.20, 0.30) (b) (0.60, 0.70) (c) (0.40, 0.50) (d) (0.10, 0.20)
Please think and answer the question based on the most likely options.

Type: Range Estimation

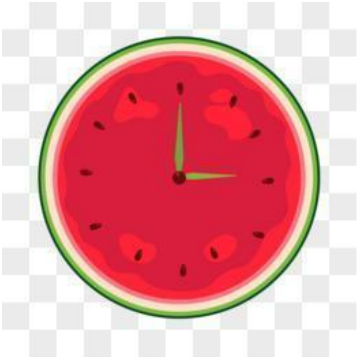
Answer:(b) (0.60, 0.70)

Gemini 2.0 Flash:

Based on the image, the value of class A appears to be approximately 0.62. Therefore, none of the options are correct.

⊗

Synthetic (Scale)



Prompt:
Question: Which of the following options is a reasonable estimate of the time the clock in the figure is pointing to?
Options:(a) 12 hours and 37 minutes (b) 10 hours and 07 minutes (c) 3 hours and 00 minutes (d) 7 hours and 22 minutes
Please think and answer the question based on the most likely options.

Type: Value Estimation

Answer:(c) 3 hours and 00 minutes

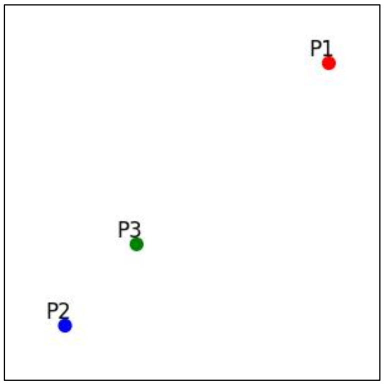
Gemini 2.0 Flash:

Based on the image, the hour hand is pointing close to the 12, and the minute hand is pointing to around the number 7. Therefore, the reasonable estimate is **(a) 12 hours and 37 minutes**.

⊗

Figure 12. Examples of VisNumBench and the results predicted by Gemini2.0 Flash (VisNumBench-Synthetic, Scale, 2/12).

Synthetic (Length)



Prompt:
Question: Which two points have the greatest distance between them?
Options: (a) P1 and P3 (b) P2 and P3 (c) P1 and P2
Please think and answer the question based on the most likely options.

Type: Value Comparison

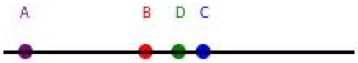
Answer:(c) P1 and P2

Gemini 2.0 Flash:

Based on the image, **(a) P1 and P3** appear to have the greatest distance between them

⊗

Synthetic (Length)



Prompt:
Question: Which of the following options is the midpoint of the line segment?
Options: (a) point A (b) point B (c) point C (d) point D
Please think and answer the question based on the most likely options.

Type: Multiplicative Estimation

Answer: `(d) point D`

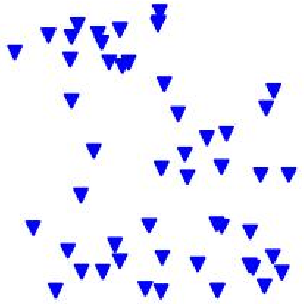
Gemini 2.0 Flash:

The midpoint of the line segment appears to be point B, as it is visually closest to the center of the line segment defined by the points shown. Therefore, the answer is (b).

⊗

Figure 13. Examples of VisNumBench and the results predicted by Gemini2.0 Flash (VisNumBench-Synthetic, Length, 3/12).

Synthetic (Quantity)



Prompt:

Question: Which is a reasonable range of estimates for the number of blue objects in the figure?

Options: (a) [1, 10) (b) [10, 20) (c) [20, 30) (d) More than 30

Please think and answer the question based on the most likely options.

Type: Range Estimation

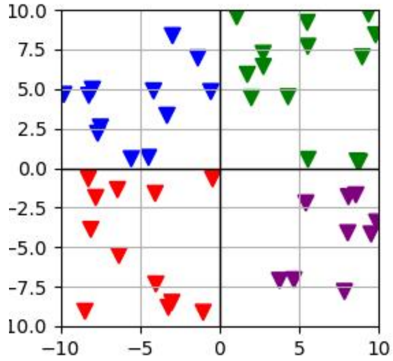
Answer:(d) More than 30

Gemini 2.0 Flash:

Based on the figure, there are more than 30 blue objects. So the answer is (d).

✓

Synthetic (Quantity)



Prompt:

Question: Which of the following colors of objects in the figure is most likely to have the highest quantity?

Options:(a) blue (b) green (c) purple (d) red

Please think and answer the question based on the most likely options.

Type: Value Comparison

Answer: (b) green

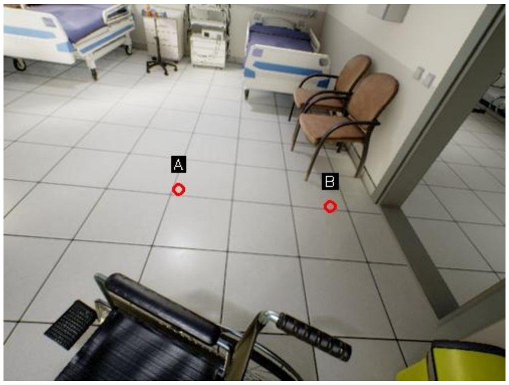
Gemini 2.0 Flash:

Based on the figure, the color with the highest quantity of objects is **blue**.Therefore, the answer is (a).

⊗

Figure 14. Examples of VisNumBench and the results predicted by Gemini2.0 Flash (VisNumBench-Synthetic, Quantity, 4/12).

Synthetic (Depth)



Prompt:
Question:The depth from point B to the viewer is X times than point A, Which of the following options is a reasonable estimate?
Options: (a) 2.3 (b) 1.0 (c) 3.6 (d) 5.7
Please think and answer the question based on the most likely options.

Type: Multiplicative Estimation

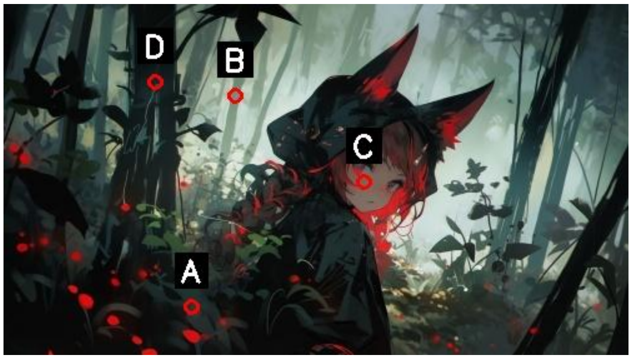
Answer:(b) 1.0

Gemini 2.0 Flash:

The depth from point B to the viewer is greater than point A. By observation, the reasonable option is (a) 2.3.

⊗

Synthetic (Depth)



Prompt:
Question: Four points are circled on the figure, labeled by A, B, C and D beside each circle. Which point appears closer to the viewer?
Options:(a) A is closer (b) B is closer (c) C is closer (d) D is closer
Please think and answer the question based on the most likely options.
Type: Value Comparison
Answer: (a) A is closer

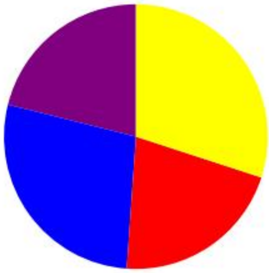
Gemini 2.0 Flash:

The point C is part of the character, who appears to be in the foreground compared to the background with A, B, and D. So, the point C is closer.Therefore, the answer is (c)

⊗

Figure 15. Examples of VisNumBench and the results predicted by Gemini2.0 Flash (VisNumBench-Synthetic, Depth, 5/12).

Synthetic (Area)



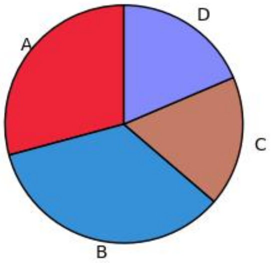
Prompt:
Question: Which of the following options is a reasonable estimate of the area percentage of the red portion of the pie chart?
Options: (a) (70%, 80%) (b) (80%, 90%) (c) (20%, 30%) (d) (40%, 50%)
Please think and answer the question based on the most likely options.
Type: Range Estimation
Answer:(c) (20%, 30%)

Gemini 2.0 Flash:

Based on the image, the red portion of the pie chart appears to be approximately one-fourth of the whole pie. One-fourth is equal to 25%. Therefore, the option that best represents this is (c) (20%, 30%).

✓

Synthetic (Area)



Prompt:
Question: In the ABCD sections of the pie chart below, which section has the largest area?
Options: (a) area A (b) area B (c) area C (d) area D
Please think and answer the question based on the most likely options.

Type: Value Comparison

Answer: (b) area B

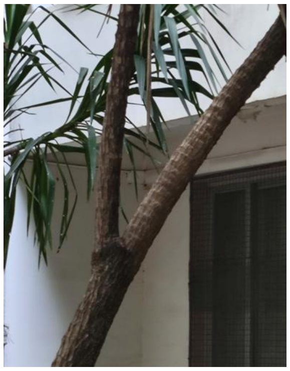
Gemini 2.0 Flash:

Based on the pie chart provided, area A appears to be the largest section.So the answer is (a) area A.

⊗

Figure 16. Examples of VisNumBench and the results predicted by Gemini2.0 Flash (VisNumBench-Synthetic, Area, 6/12).

Real (Angle)



Prompt:
Question: Which of the following options best estimates the angles between the branches in the figure?
Options: (a) (90, 100) degrees (b) (70, 80) degrees (c) (30, 40) degrees (d) (130, 140) degrees
Please think and answer the question based on the most likely options.

Type: Range Estimation

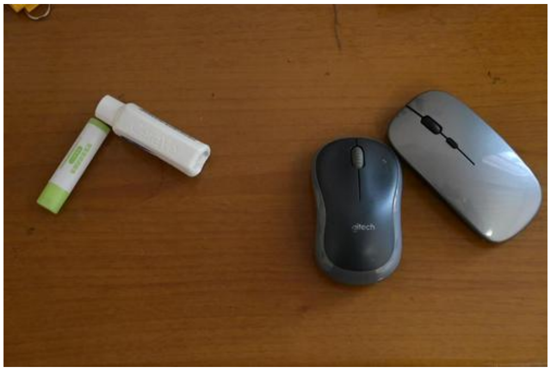
Answer: (c) (30, 40) degrees

Gemini 2.0 Flash:

Based on the image, the angles between the branches are quite wide. Option (d), (130, 140) degrees, seems to best estimate the angles.

✗

Real (Angle)



Prompt:
Question: The angle formed by the objects on the left side is X times the angle formed by the objects on the right side of the figure. Which of the following options is a reasonable estimate for X?
Options:(a) 1.0 (b) 3.3 (c) 2.0 (d) 4.2
Please think and answer the question based on the most likely options.

Type: Multiplicative Estimation

Answer: (c) 2.0

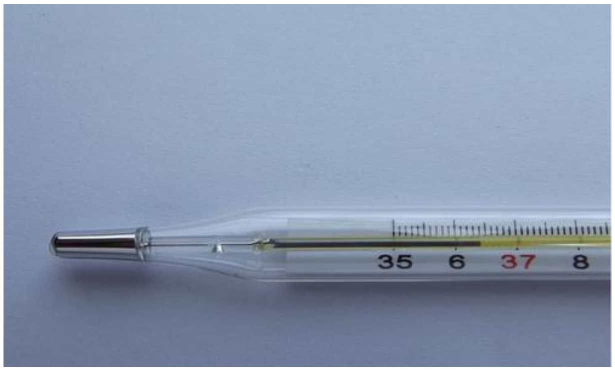
Gemini 2.0 Flash:

Let's analyze the angles formed by the objects on the left and right sides of the image.On the left, the angle between the two tubes is approximately 60 degrees.On the right, the angle between the two mice is approximately 20 degrees.To find the value of X, we divide the angle on the left by the angle on the right:X = 60 degrees / 20 degrees = 3Therefore, the angle formed by the objects on the left side is approximately 3 times the angle formed by the objects on the right side.Based on the options provided:(a) 1.0(b) 3.3(c) 2.0(d) 4.2The most reasonable estimate for X is 3.3.Final Answer: The final answer is $\boxed{3.3}$

✗

Figure 17. Examples of VisNumBench and the results predicted by Gemini2.0 Flash (VisNumBench-Real, Angle, 7/12).

Real (Scale)



Prompt:
Question: Which of the following options is an reasonable estimate of the temperature in degrees Celsius that the thermometer in the figure is pointing to?
Options: (a) 36.4 (b) 37.1 (c) 38.0 (d) 39.4
Please think and answer the question based on the most likely options.

Type: Value Estimation

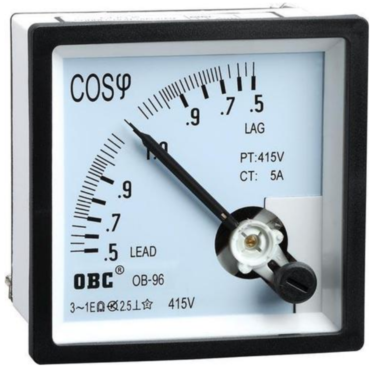
Answer: (a) 36.4

Gemini 2.0 Flash:

Based on the image, the thermometer reading appears to be slightly above 37 degrees Celsius. Therefore, the most reasonable estimate from the options is (b) 37.1.

⊗

Real (Scale)



Prompt:
Question: Which of the following options is an reasonable estimate of the value that the pointer in the figure is pointing to, regardless of whether it is positive or negative?
Options:(a) 8 (b) 1 (c) 10 (d) 3
Please think and answer the question based on the most likely options.

Type: Value Estimation

Answer: (c) 10

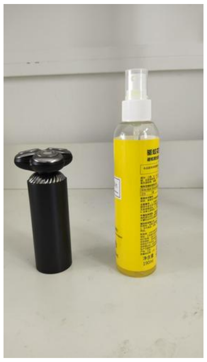
Gemini 2.0 Flash:

The pointer is pointing very close to the 1.0 mark.The question asks for a "reasonable estimate of the value that the pointer... is pointing to".Of the provided options, the value closest to 1 is (b) 1.Therefore, the answer is (b).

⊗

Figure 18. Examples of VisNumBench and the results predicted by Gemini2.0 Flash (VisNumBench-Real, Scale, 8/12).

Real (Length)



Prompt:
Question: The right object is X times higher than left object. Which of the following options is a reasonable estimate for X?
Options: (a) 2.8 (b) 1.5 (c) 3.7 (d) 4.7
Please think and answer the question based on the most likely options.

Type: Multiplicative Estimation

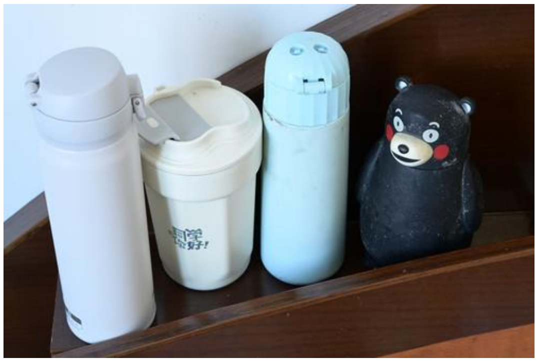
Answer: (b) 1.5

Gemini 2.0 Flash:

Based on visual estimation, the right object (spray bottle) appears to be roughly 2.5 to 3 times higher than the left object (electric razor).Therefore, option (a) 2.8 is the most reasonable estimate.

⊗

Real (Length)



Prompt:
Question: The objects from left to right in the figure are object A, object B, object C, and object D. Which object has the largest length?
Options:(a) object A (b) object B (c) object C (d) object D
Please think and answer the question based on the most likely options.

Type: Value Comparison

Answer: (a) object A

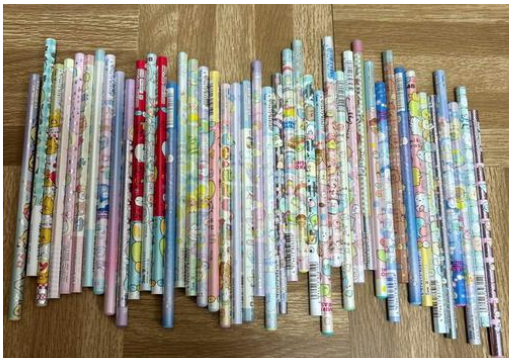
Gemini 2.0 Flash:

Based on the image, object A appears to have the largest length.The answer is (a).

✓

Figure 19. Examples of VisNumBench and the results predicted by Gemini2.0 Flash (VisNumBench-Real, Length, 9/12).

Real (Quantity)



Prompt:
Question: Which of the following options is a reasonable estimate of pencil in the figure?
Options: (a) 70 (b) 26 (c) 43 (d) 54
Please think and answer the question based on the most likely options.

Type: Value Estimation

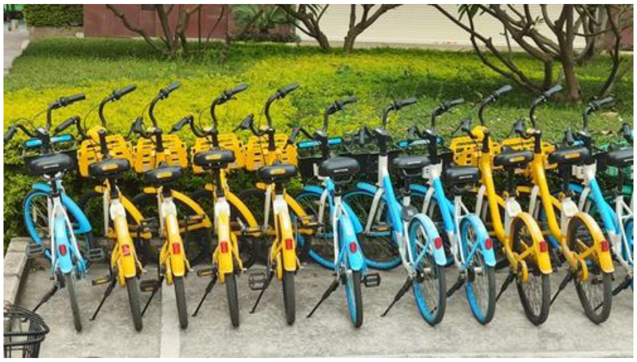
Answer:(c) 43

Gemini 2.0 Flash:

Based on the image, a reasonable estimate of the number of pencils is 54. So the answer is (d)

⊗

Real (Quantity)



Prompt:
Question: Which of the following options is a reasonable range of estimates for the number of bikes in the figure?
Options:(a) [1, 10) (b) [10, 30) (c) More than 30
Please think and answer the question based on the most likely options.

Type: Range Estimation

Answer: (b) [10, 30)

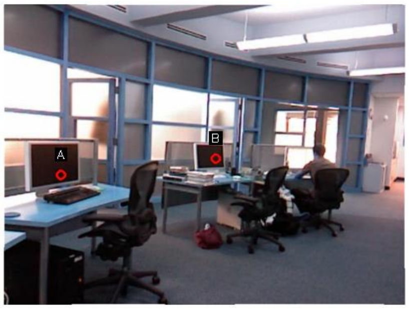
Gemini 2.0 Flash:

Based on the image, there are clearly more than 10 bikes visible. Counting them suggests there are more than 30 bikes.Therefore, the answer is (c).

⊗

Figure 20. Examples of VisNumBench and the results predicted by Gemini2.0 Flash (VisNumBench-Real, Quantity, 10/12).

Real (Depth)



Prompt:
Question: The distance from point B to the viewer is X times than point A, Which of the following options is a reasonable estimate for X?
Options: (a) 2.0 (b) 3.1 (c) 5.3 (d) 6.7
Please think and answer the question based on the most likely options.

Type: Multiplicative Estimation

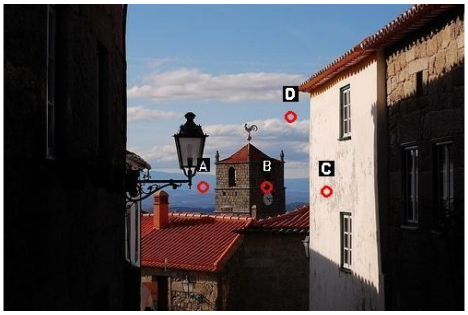
Answer: (a) 2.0

Gemini 2.0 Flash:

Based on the image, point B looks to be approximately 3 times farther away than point A.So the answer is (b) 3.1

⊗

Real (Depth)



Prompt:
Question: Four points are circled on the figure, labeled by A, B, C and D beside each circle. Which of the following options is most likely to be the order in which these points are located from closest to farthest from the viewer?
Options:(a) ACDB (b) BCAD (c) ACBD (d) CBAD
Please think and answer the question based on the most likely options.

Type: Value Comparison

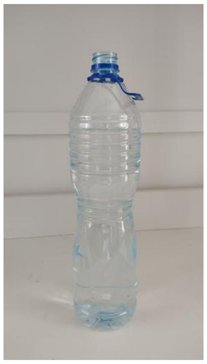
Answer: (d) CBAD

Gemini 2.0 Flash:
Okay, let's analyze the image to determine the most likely order of the points from closest to farthest.* **C:** This point is on the building that is on the right-hand side of the image. The building is close to the viewer. Therefore, it's likely that C is the closest point to the viewer.* **A:** This point is on the lamppost which is closer than the buildings at the back.* **B:** This point is on the bell tower further away from the building on the right. It appears farther than the right building with point C.* **D:** This point is near the upper part of the building where point C is located. However, given its position in the image and its relative height, it is more likely to be in the far background, behind the B.Therefore, the most likely order from closest to farthest is C, A, B, D, which corresponds to option **(c) ACBD**.

⊗

Figure 21. Examples of VisNumBench and the results predicted by Gemini2.0 Flash (VisNumBench-Real, Depth, 11/12).

Real (Volume)



Prompt:
Question: The total capacity of the bottle is 1500 ml. Which of the following options is a reasonable estimate of the volume of water in the bottle?
Options: (a) (350, 450) ml (b) (1305, 1405) ml (c) (1450, 1550) ml (d) (1050, 1150) ml
Please think and answer the question based on the most likely options.

Type: Range Estimation

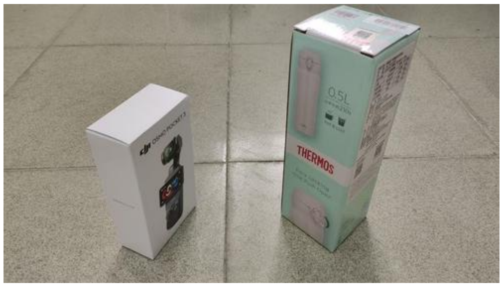
Answer: (b) (1305, 1405) ml

Gemini 2.0 Flash:

Based on the image, the water bottle appears to be about 25-30% full. Given a total capacity of 1500 ml, the volume of water is likely between 375 ml and 450 ml. Therefore, the most reasonable estimate is (a) (350, 450) ml.

⊗

Real (Volume)



Prompt:
Question: The volume of the right object is X times greater than that of the left object. Which of the following options is a reasonable estimate for X?
Options:(a) 3.1 (b) 5.2 (c) 2.0 (d) 6.6
Please think and answer the question based on the most likely options.

Type: Multiplicative Estimation

Answer: (c) 2.0

Gemini 2.0 Flash:

Based on visual estimation, the right object (the Thermos box) appears to be roughly 5 times larger in volume than the left object (the DJI Osmo Pocket 3 box).Therefore, the answer is (b) 5.2.

⊗

Figure 22. Examples of VisNumBench and the results predicted by Gemini2.0 Flash (VisNumBench-Real, Volume, 12/12).