SITE: towards Spatial Intelligence Thorough Evaluation

Supplementary Material

I. Datasets Surveyed

As mentioned in the main text, we have collected a total of 31 computer vision datasets, comprising 22 image-based datasets, 8 video-based datasets and 1 newly annotated dataset.

Image-based datasets. We have collected the samples from the validation/test split of the following 22 image-based datasets:

- Blink [20]: a comprehensive benchmark designed to evaluate multimodal large language models(MLLMs) across broad visual perception tasks.
- CLEVR [31]: a visual question answering dataset containing various aspects of visual reasoning tasks.
- CVBench [66]: a vision-centric benchmark evaluating 2D and 3D understanding of the models.
- GQA [29]: a visual question answering benchmark designed to test compositional reasoning and spatial understanding constructed from structured scene representations.
- IconQA [48]: a dataset targeting diagram-based question answering, challenging models to interpret and reason over abstract visual representations.
- LogicVista [74]: a benchmark aimed at assessing the logical reasoning capabilities of MLLMs through structured visual tasks.
- MMBench [46]: a large-scale benchmark for evaluating the performance of multimodal models across a wide range of vision-language tasks.
- MME [18]: a comprehensive evaluation benchmark for MLLMs, covering various aspects of perception and cognition abilities.
- MME-RealWorld [82]: a real-world multimodal evaluation benchmark that tests models on practical perception tasks.
- MMIU [55]: a comprehensive benchmark designed to evaluate multi-image tasks on MLLMs.
- MMTBench [77]: an evaluation benchmark containing massive multimodal tasks from various scenarios.
- MMVet [80]: an evaluation suite focusing on the integration of multiple Vision-Language capabilities.
- MMVP [67]: an MLLM evaluation benchmark focusing on CLIP-blind pairs.
- MuirBench [70]: a comprehensive benchmark targeting robust multi-image understanding abilities of multimodal models.
- SAT [60]: a spatial training dataset with static and dynamic spatial reasoning tasks.
- SeedBench [37–39]: an evaluation benchmark for generative comprehension capabilities of MLLMs.
- SpatialEval [71]: a novel benchmark that covers different aspects of spatial reasoning in textual and visual formats.
- SPEC [58]: a synthetic dataset designed to test the fine-grained vision-language understanding of models.
- VQAv2 [36]: an enhanced version of the Visual Question Answering dataset, providing more balanced question-answer pairs to reduce language biases and better evaluate visual understanding.
- VSR [45]: a benchmark designed to assess visual spatial reasoning capabilities within images.
- VStarBench [72]: a visual question answering benchmark that focuses on detailed visual grounding on high-resolution images.
- 3DSRBench [49]: a comprehensive 3D spatial reasoning benchmark on diverse entities.

Video-based QA datasets. Our video-based QA samples are collected from the following 8 video-based datasets:

- ActivityNetQA [81]: a large-scale video question answering dataset based on ActivityNet, designed to evaluate models' abilities to comprehend and reason about complex human activities in videos.
- MLVU [83]: a comprehensive benchmark designed for long video understanding
- MVBench [42]: a comprehensive benchmark for multimodal video understanding, assessing models on a variety of tasks.
- Open-EQA [50]: an open-ended embodied question answering dataset that tests models' abilities to interact with and reason about 3D environments through videos and natural language queries.
- TGIFQA [30]: a dataset for spatiotemporal reasoning in video question answering through tasks like action recognition and repetition counting.
- TVQA [36]: a video question answering dataset constructed from TV shows, focusing on temporal and contextual reasoning.
- VideoMME [19]: a comprehensive benchmark for evaluating multimodal large language models on video understanding tasks.
- VSI-Bench [75]: a benchmark designed to evaluate spatial reasoning abilities of MLLMs, focusing on tasks that require understanding spatial relationships within indoor scenes.

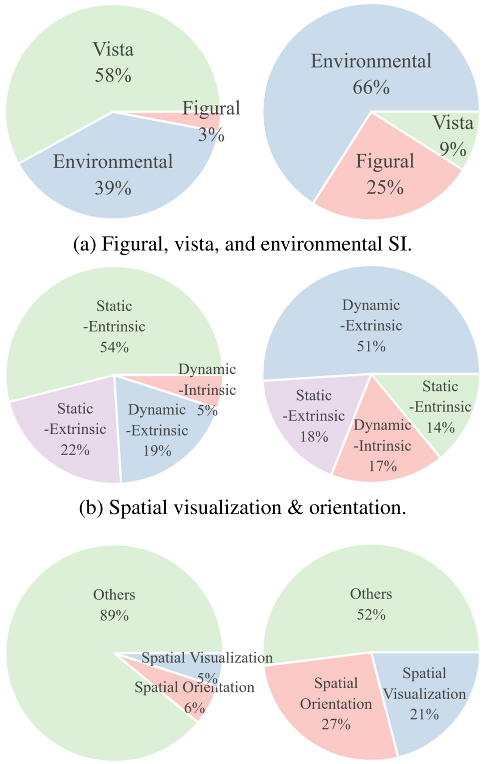
Ego-Exo4D [23]: a large-scale, multimodal, multiview video dataset capturing skilled human activities from synchronized egocentric and exocentric perspectives. It encompasses over 1,286 hours of video data collected from

740 participants across 13 cities, featuring diverse tasks such as cooking, sports, and music. The dataset includes rich annotations like expert commentary, narrate-and-act descriptions, and atomic action labels, supporting benchmarks in fine-grained activity recognition, proficiency estimation, cross-view translation, and pose estimation.

II. Category Filtering

After we collected all the data and their text annotation information (questions, options, answers, descriptions, prompts), we used GPT-4o and manual inspection to derive 6 coarse level classifications of spatial intelligence for the data with category labels; for the data without category labels, we carefully designed the following prompts and used GPT-4o to perform few-shot classification of the text annotation information, shown as Figure 9, 10 and 11. The six categories of spatial intelligence are defined and described as follows:

1. **Counting & Existence.** Evaluates the model's ability to detect and quantify object occurrences within static images or video sequences. This includes recognizing the presence or absence of specific objects and accurately counting their instances across frames.
2. **Spatial Relationship Reasoning.** Assesses the model's capacity to infer relative spatial relationships between objects. This encompasses understanding positional attributes such as proximity, occlusion, containment, and directional relations (e.g., left/right, above/below).
3. **Multi-View Reasoning.** Measures the model's ability to integrate and interpret information across multiple viewpoints. This includes understanding object appearances from different perspectives, reasoning about occluded or unseen parts, and reconstructing spatial arrangements from limited observations.
4. **3D Information Understanding.** Evaluates the model's capability to perceive and represent three-dimensional object properties. This involves recognizing shape, depth, surface structure, and spatial extent, as well as reasoning about object interactions in a 3D environment.
5. **Object Localization & Positioning.** Tests the model's accuracy in determining object locations within an image or scene. This includes detecting precise spatial coordinates, generating bounding boxes or keypoints, and performing spatial alignment.
6. **Movement Prediction & Navigation.** Assesses the model's ability to predict object motion and infer navigational paths within dynamic environments. This includes trajectory forecasting, motion intent recognition, and decision-making based on spatial and temporal cues.



(a) Figural, vista, and environmental SI.

(b) Spatial visualization & orientation.

(c) 2x2 classification.

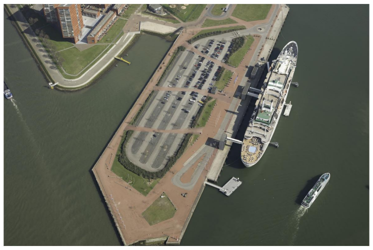
Figure 4. Category distribution by cognitive classification systems. **Left:** distribution before balancing. **Right:** our final benchmark's distribution.

III. Dataset Statistics

Besides, we include more dataset statistics, a radar chart comparing various models, more examples in our dataset, an example about the frame-reordering task, and a bar chart to reveal data decomposition.

Correlation Analysis. We collect performance scores from four vision-language models—Qwen2.5-VL and InternVL-2.5 series—across various benchmarks, along with their corresponding performances on LIBERO-Spatial when used as VLA backbones. We compute the Pearson correlation coefficient to measure the linear relationship between benchmark performance and LIBERO-Spatial results, as shown in Table 6. The analysis reveals that our SITE benchmark exhibits one of the strongest positive correlations, indicating that spatial intelligence plays a critical role in robotic manipulation tasks compared to general VQA capabilities (e.g., RealWorldQA, Q-Bench), OCR capabilities (OCRBench), scientific knowledge (ScienceQA), and object probing (POPE). Notably, MathVista also demonstrates a high positive correlation. We hypothesize that this may be attributed to the role of reasoning ability in enhancing performance on embodied tasks.

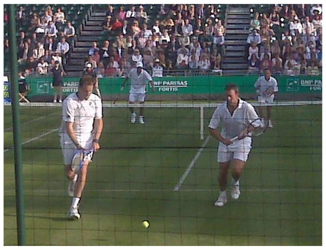
Counting & Existence



Question: How many ships are there on the water surface?

Options:
(A) 4
(B) 5
(C) 3
(D) 2
(E) This image doesn't feature the count

Answer: (C) 3

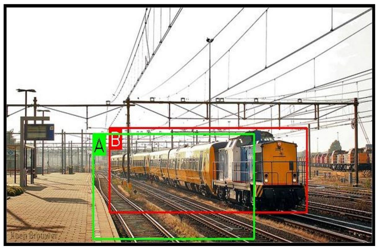


Question: How many people are playing in tennis?

Options:
(A) 4
(B) 5
(C) 6
(D) 3

Answer: (A) 4

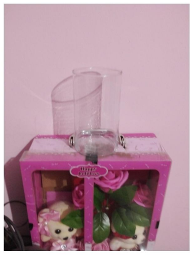
Object Localization



Question: Which bounding box more accurately localizes and encloses the train (railroad vehicle)?

Options:
(A) Box A
(B) Box B

Answer: (B) Box B

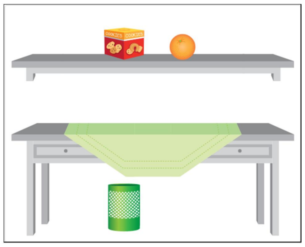


Question: Please detect all transparent foreground instances in this image. Please provide the bounding box coordinates for the described object or area using the format [x1, y1, x2, y2].

Options:
(A) [0.364, 0.245, 0.663, 0.601]
(B) [0.739, 0.618, 0.975, 0.97]
(C) [0.364, 0.245, 0.645, 0.625]
(D) [0.364, 0.245, 0.662, 0.61]

Answer: (C) [0.364, 0.245, 0.645, 0.625]

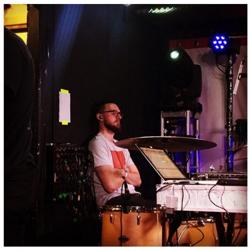
Spatial Relationship



Question: Which object is next to the box of cookies?

Options:
(A) orange
(B) table cover
(C) cylinder

Answer: (A) orange

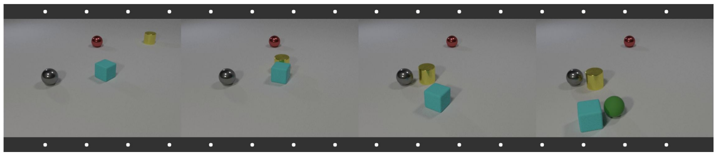


Question: What is the position of the man relative to the drum set?

Options:
(A) In front of
(B) To the left of
(C) To the right of
(D) Behind

Answer: (D) Behind

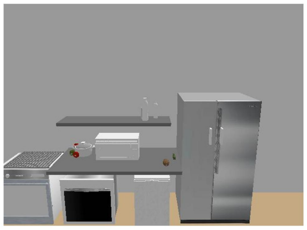
Movement and Navigation



Question: In which direction does the yellow cylinder move?

Options:
(A) Down and to the left (B) Up and to the left
(C) The object is stationary (D) Up and to the right

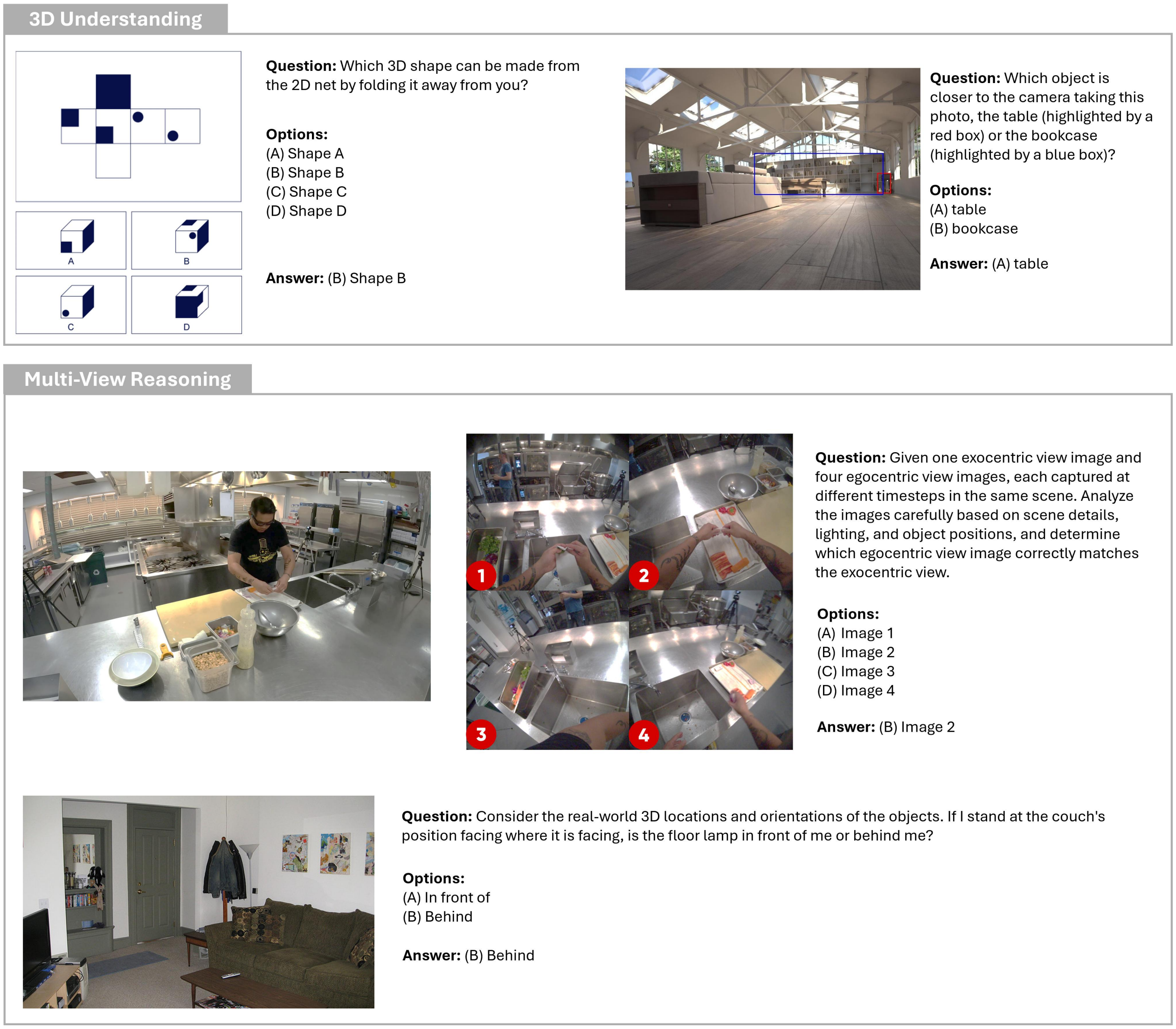
Answer: (A) Down and to the left



Question: Please generate detailed steps to complete the following task: put the bottle on the fridge.

Options:
(A) (1) Reach the bottle. (2) Grab it and lift it up.(3) Put it to the proper position.
(B) (1) Get the bottle from the fridge. (2) Drink from the bottle. (3) Leave the bottle on the table.
(C) (1) Open the fridge. (2) Take out the bottle. (3) Close the fridge.
(D) (1) Find the bottle in the fridge. (2) Pour the liquid from the bottle. (3) Put the bottle back in the fridge.
Answer: (A) (1) Reach the bottle. (2) Grab it and lift it up.(3) Put it to the proper position.

Figure 5. Two samples for each spatial category. (Part I)



3D Understanding

Question: Which 3D shape can be made from the 2D net by folding it away from you?

Options:
(A) Shape A
(B) Shape B
(C) Shape C
(D) Shape D

Answer: (B) Shape B

Question: Which object is closer to the camera taking this photo, the table (highlighted by a red box) or the bookcase (highlighted by a blue box)?

Options:
(A) table
(B) bookcase

Answer: (A) table

Multi-View Reasoning

Question: Given one exocentric view image and four egocentric view images, each captured at different timesteps in the same scene. Analyze the images carefully based on scene details, lighting, and object positions, and determine which egocentric view image correctly matches the exocentric view.

Options:
(A) Image 1
(B) Image 2
(C) Image 3
(D) Image 4

Answer: (B) Image 2

Question: Consider the real-world 3D locations and orientations of the objects. If I stand at the couch's position facing where it is facing, is the floor lamp in front of me or behind me?

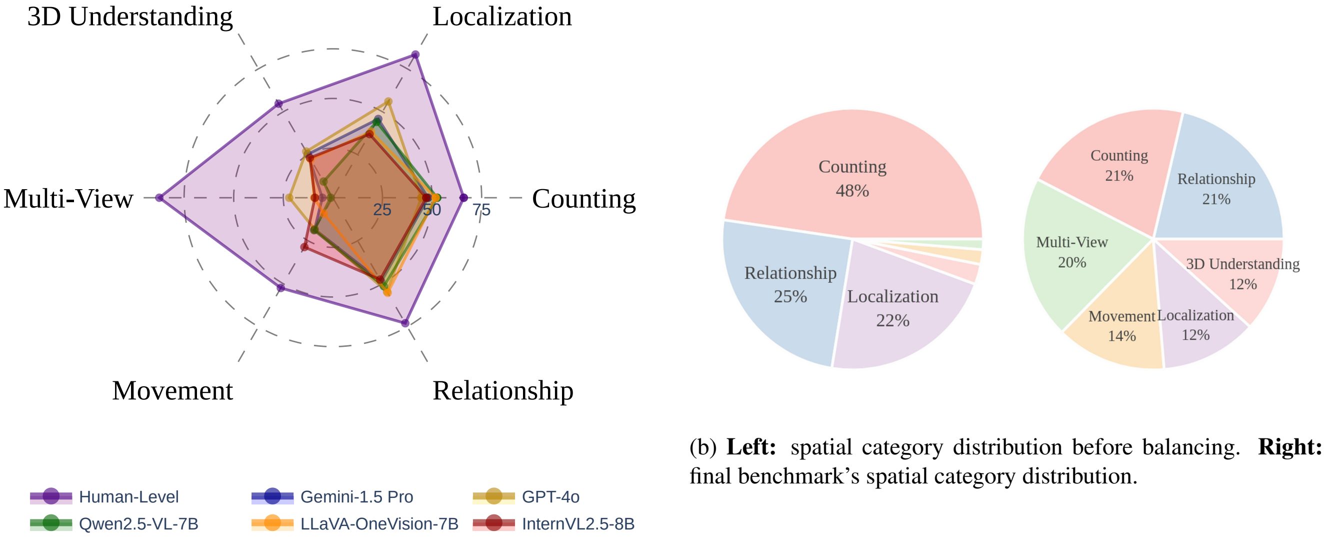
Options:
(A) In front of
(B) Behind

Answer: (B) Behind

Figure 6. Two samples for each spatial category. (Part II)

Model	L2 Dist ↓	Sim SR (%) ↑	MathVista ↑	POPE ↑	ScienceQA ↑	OCRBench ↑	RealWorldQA ↑	QBench ↑	SITE ↑
LLaVA-OV-0.5B	0.268 ± 0.241	0.0	35.9	87.8	67.5	58.3	51.8	62.5	18.4
LLaVA-OV-7B	0.142 ± 0.172	0.0	62.6	88.4	95.4	62.2	69.9	78.9	30.2
Qwen2.5-VL-3B	0.139 ± 0.153	0.0	61.2	85.9	81.4	82.8	65.5	74.9	29.5
Qwen2.5-VL-7B	0.030 ± 0.040	38.0	68.1	85.9	89.0	88.8	68.4	77.7	31.4
Correlation	-	-	**0.935**	-0.602	0.749	0.832	0.842	0.847	**0.902**

Table 6. **Correlation between SI and robotics manipulation on Libero Spatial.** The Correlation row shows the Pearson correlation coefficient between the negated mean L2 distance and different benchmark scores. The **bold** numbers show a higher correlation.



(b) **Left:** spatial category distribution before balancing. **Right:** final benchmark's spatial category distribution.

(a) Different models' performance in a glance.

Figure 7. Data distribution and model's performance under six coarse spatial categories.

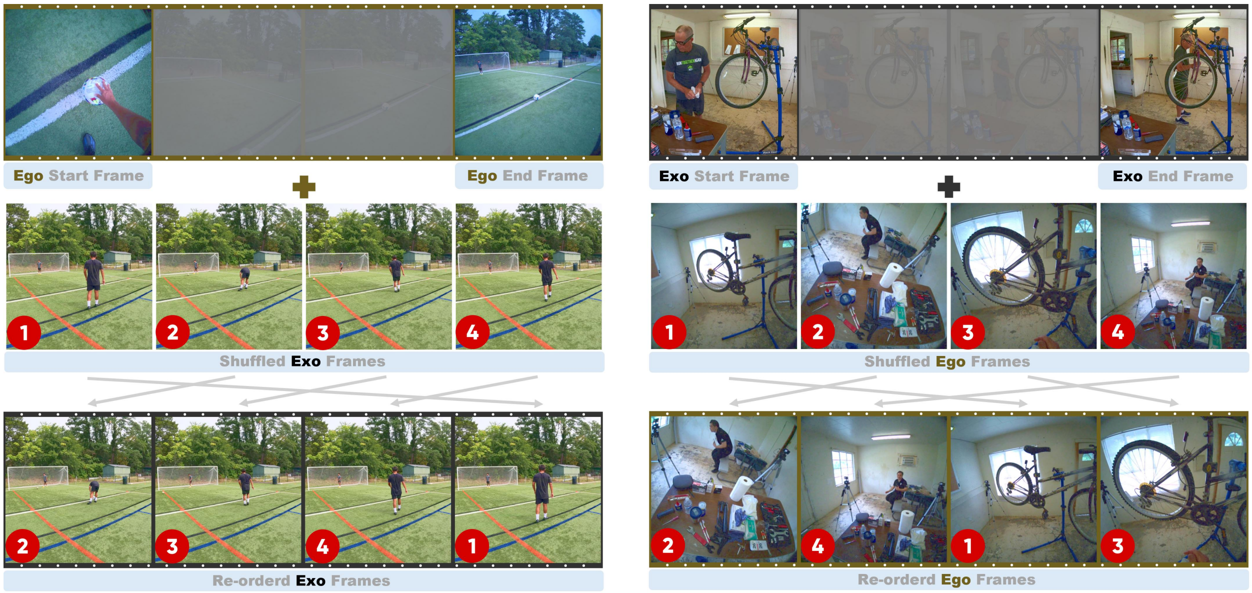


Figure 8. **Ego-Exo frames reordering tasks.** Given the start and end frames of a video clip in an egocentric view, and four randomly shuffled frames from the same clip in exocentric views, the model is tasked with reordering the four shuffled frames into their correct temporal sequence (or vice versa).

```
enhanced_prompt_in_english = """
System Role Instructions (System):
You are ChatGPT, a large language model trained by OpenAI. Your goal is to help the user
    classify a series of Q&A pairs to determine whether they are spatially related. If a
    pair is indeed spatially related, you must further categorize it into one of the
    specified categories.

You must follow these rules:
1. If the Q&A content is NOT related to spatial relationships, simply answer:
   No
2. If the Q&A content IS related to spatial relationships, answer:
   Yes. This is a X problem because ...
   where X must be chosen from the following list:
   - Counting & Existence
   - Object Localization & Positioning
   - Spatial Relationship Reasoning
   - Depth & 3D Understanding
   - Multi-view & Cross-Image Reasoning
   - Movement Navigation & Intent Prediction
   - Other spatial category can not be sure.
3. If multiple Q&A pairs (N Q&A pairs) are provided in a single input, you must apply the
    same classification steps to each Q&A pair in the order they appear, and output the
    result for each pair in that order.

User Role Instructions (User):
Below are examples and their reference outputs (few-shot examples). Please study the logic
    and answer format shown in these examples before performing the classification:

[Example 1]
Input:
Question: How many blue floats are there?
Select from the following choices.
(A) 0
(B) 3
(C) 2
(D) 1
Answer: (D) 1

Output:
Yes. This is a Counting & Existence problem because it asks about the number of objects.

[Example 2]
Input:
Question: What is the position of the catcher relative to the home plate?
Options:
A: The catcher is to the left of the home plate.
B: The catcher is to the right of the home plate.
C: The catcher is behind the home plate.
D: The catcher is in front of the home plate.
Answer: A: The catcher is to the left of the home plate.

Output:
Yes. This is a Spatial Relationship Reasoning problem because it asks about the relative
    relation between two objects.


[Example 3]
Input:
Question: Where is the bongo?
Answer: On top of the brown shelf.

Output:
Yes. This is an Object Localization & Positioning problem because it asks about the location
    of an object.
"""
```

Figure 9. Few-shot prompt used for spatial category classification. (Part I)

```
enhanced_prompt_in_english += """
[Example 4]
Input:
Question: Here are some images and their corresponding depth images: <img><img><img><img>.
Please select the correct corresponding image for the target image: <img>.
The option images are: <img><img><img><img>
Answer: The second image.

Output:
Yes. This is a Depth & 3D Understanding problem because it asks about depth information.


[Example 5]
Input:
Question: These images are frames from a video. The video shows a static scene, and the
    camera is either moving clockwise (left) or counterclockwise (right) around the object.
The first image is from the beginning of the video, and the second image is from the end. Is
    the camera moving left or right during the filming?
Select from the following options:
(A) left
(B) right
Answer: (A) left

Output:
Yes. This is a Multi-view & Cross-Image Reasoning problem because it focuses on multi-view
    information of the object and determines the camera’s rotation direction.


[Example 6]
Input:
Question: This is a navigation video of an agent following the instruction: "Exit the kitchen
     and wait in the sitting room, near the loveseat."
What is the next action it should take?
Options: Move forward. / Turn right and move forward. / Turn left and move forward. / Stop.
Answer: Stop

Output:
Yes. This is a Movement Navigation & Intent Prediction problem because it asks about the next
     action of the agent.


[Example 7]
Input:
Question: What is the color of the cat?
Answer: The cat is black.

Output:
No

[Example 8]
Input:
Question: Please correctly describe this set of images from a spatial context perspective.
Select from the following choices:
A: There is a box with four items, and three of them are touching the side.
B: There is a box with five items, all in the center.
C: There is a box with three items, and four of them are touching the side.
D: There is a bag with four items, and three of them are touching the side.
Answer: A.

Output:
Yes, but it’s hard to determine the category. This is a spatially related problem because it
    asks about the spatial context of the objects.
(If you are unsure which exact category it belongs to, choose "Other spatial categories can
    not be sure.")
"""
```

Figure 10. Few-shot prompt used for spatial category classification. (Part II)

```
enhanced_prompt_in_english += """
[Main Task]
1. Read the new Q&A input(s).
2. First, decide whether each Q&A is related to spatial relationships.
3. If NOT related, simply output:
   No
4. If related, output:
   Yes. This is a [specific category] problem because [reason].
   where [specific category] is strictly from the list:
   - Counting & Existence
   - Object Localization & Positioning
   - Spatial Relationship Reasoning
   - Depth & 3D Understanding
   - Multi-view & Cross-Image Reasoning
   - Movement Navigation & Intent Prediction
   - Other spatial categories can not be sure.
5. If multiple Q&A pairs are given together (N Q&A pairs), repeat steps 2 to 4 for each Q&A
    pair in order, returning the results in the same order and prefixing each result with an
     index '1. ', '2. ', etc.

Please keep the output style consistent and follow all the rules above.
"""
```

Figure 11. Few-shot prompt used for spatial category classification. (Part III)

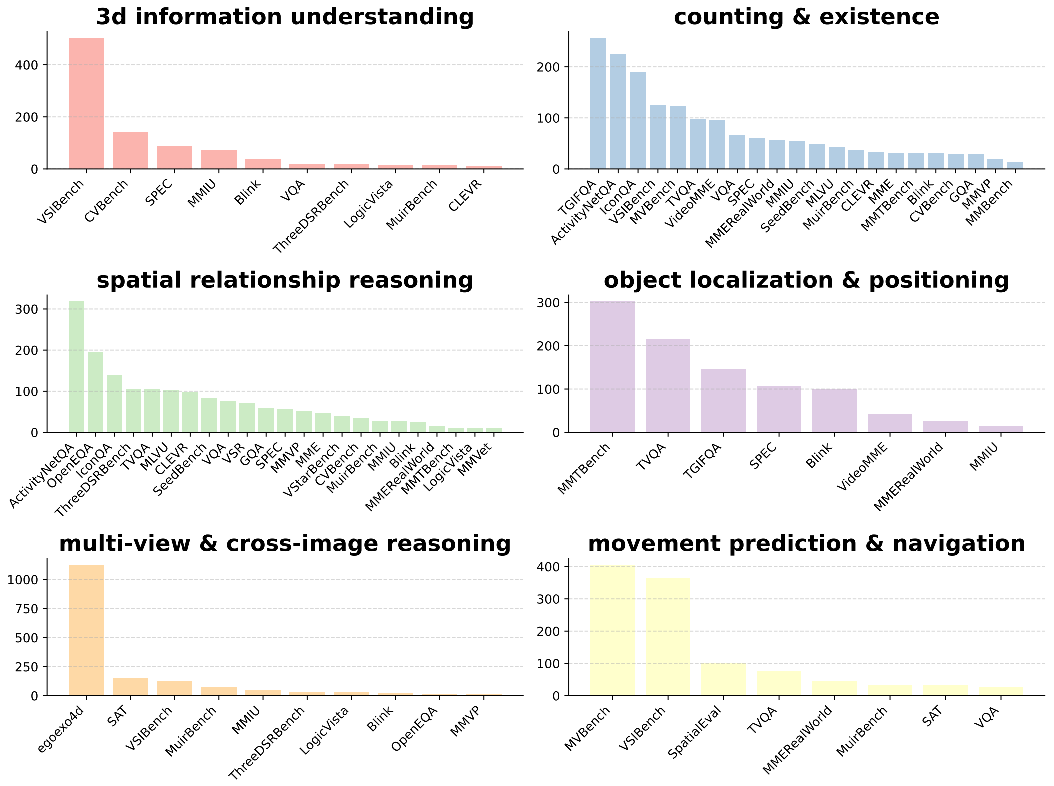


Figure 12. Dataset composition for each spatial category.