

VLM4D: Towards Spatiotemporal Awareness in Vision Language Models

Supplementary Material

Appendix Outline

- Section A: Statistics of VLM4D benchmark
- Section B: Evaluation setup
- Section C: Analysis of Existing Video Instruction Tuning Datasets
- Section D: Examples of VLM’s response with Chain of Thought and Direct Output prompting.
- Section E: More details of 4D reconstruction using Feature4X.

A. VLM4D Benchmark Statistics

Dataset Statistics	Video	QA pair
Real Samples	600	1,371
Synthetic Samples	400	445
Total Samples	1,000	1,816

Table A. VLM4D Dataset Breakdown

Tab. A presents the breakdown of our VLM4D benchmark dataset. Additionally, Fig. A visualizes the detailed performance of VLMs across different question categories. For models that support only image input, we convert videos into multi-frame image sequences, using the maximum number of frames allowed within the model’s context window. For models that support video input, we follow their default frame rate settings, typically 1 fps.

B. Evaluation setup

Task	GPU Configuration
Model Evaluation	8xA100
4D Feature Field Reconstruction	1xA100

Table B. Evaluation Type and GPU Requirements

C. Video Instruction Tuning Dataset Analysis

We begin by analyzing four individual collections of datasets, which together contribute a substantial body of data for our experiments. The datasets used in this study are:

- ShareGPT-4o
- VideoChat2-IT
- ShareGPT4Video
- LLaVA-178k

As shown in Tab. C, these datasets contain over 2 million samples in total, a robust foundation for evaluating and

benchmarking the spatiotemporal validity of the highly used video instruction tuning datasets.

We show our target strings in Tab. E. In our comprehensive analysis of the ShareGPT-4o dataset (Fig. E), we observed that over 40% of the captions incorporate at least one target category. As depicted in Fig. G, our target string search highlights a significant emphasis on the directional descriptors “left” and “right.” Furthermore, the analysis of negative samples, illustrated in Fig. F, indicates that these directional terms are seldom employed in conjunction with rotational or translational actions. This observation is further substantiated by the minimal overlaps between directional descriptors and action-related terms, as shown in Fig. G. Collectively, this shows a notable gap in the dataset’s ability to capture complex spatiotemporal relationships, particularly those involving dynamic textures and nuanced motion patterns. The statistics are summarized in Tab. D.

Dataset	Count
VideoChat2-IT	423,497
ShareGPT4Video	40,178
LLaVA-178K	1,627,017
ShareGPT-4o	2,111
Total	2,092,803

Table C. No. of Video Instruction Tuning Samples (QA pair)

Caption Category	Count	Percentage (%)
With Target String	1,164,006	55.6
Without Target String	928,897	44.4

Table D. Distribution of Captions Containing Target Strings

D. Examples of VLM’s response

Please refer to Fig. I - M for detailed responses from all evaluated VLMs under both chain-of-thought (CoT) and direct output (DO) prompting, based on the given example video and question.

E. Details of 4D Reconstruction

We utilize the Feature4X framework (Fig. H) for 4D reconstruction experiments conducted on our dataset. Given an input monocular RGB video, Feature4X reconstructs the

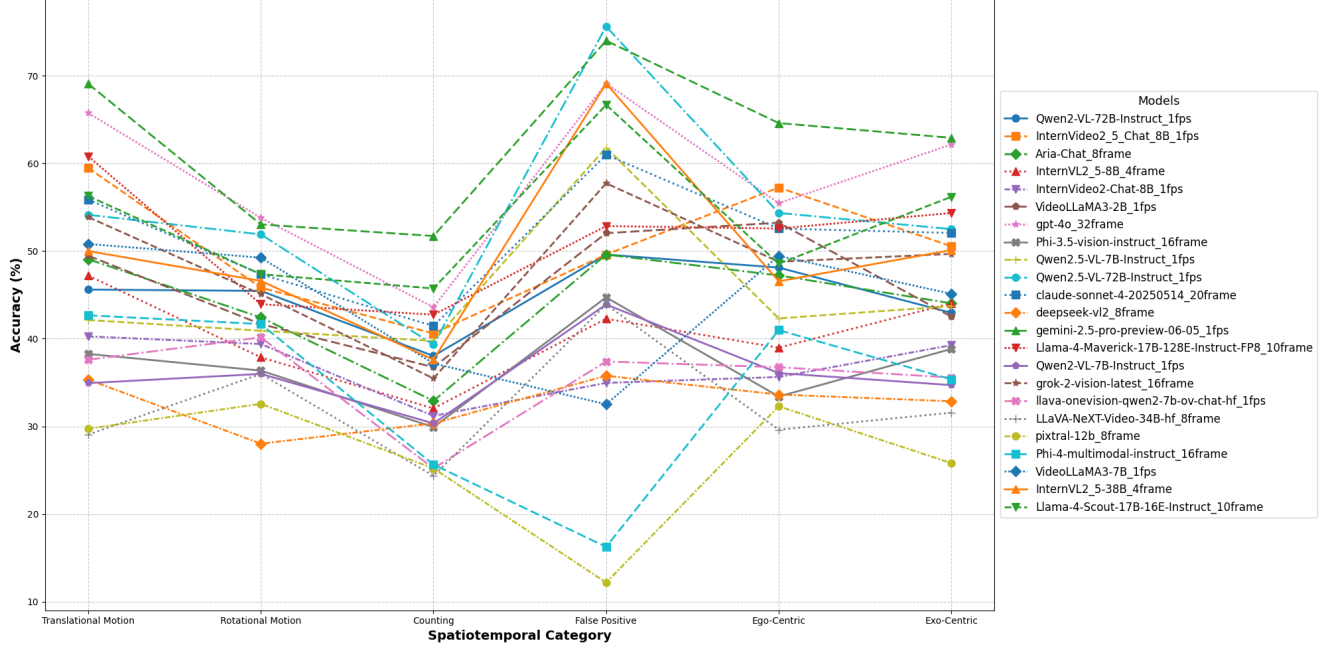


Figure A. Performance comparison of various VLMs across annotated question categories including counting, rotational motion, translational motion, false positives, and action recognition

Category	Terms
Directional Descriptors	left, right, up, down, north, south, east, west, ahead, behind, towards the front, away from the, to the left, toward, to the right, in front of, behind you, side to side, straight ahead, high ground, low ground, left and right, front and back, top and bottom, northern, southern, eastern, western, northeast, northwest, southeast, southwest, around
Translational Actions	moving, running, walking, sprinting, gliding, sliding, crawling, trotting, jogging, skipping, bounding, rushing, hurrying, traveling, shifting, advancing, progressing, traversing, racing, zooming, going fast, going
Rotational Descriptors	rotate, revolve, spin, gyrate, twirl, whirl, twist, turn, pivot, flip, roll, spiral, swing, shake, oscillate, swing around, rolls, roll
Rotational Actions	clockwise, anticlockwise, turn right, turn left, spin, rotate, revolve, twist, pivot, gyrate, whirl, rotating, spinning, turning 360 degrees, turning 180 degrees, rotation, twisting, turning around, circular motion, turning 90 degrees,
Perspective Descriptors	camera's perspective, frame's perspective, viewpoint, point of view, line of sight, from above, from below, from the side, from the front, from behind, top view, bottom view, rear view, side view, front view, bird's eye view, aerial view, close-up view, distant view

Table E. Video Instruction Tuning Datasets Categories & Target strings Analysis

dynamic 3D scene by employing dynamic 3D Gaussians, specifically Dynamic 3D Gaussian Splatting, which represent dynamic foreground elements that deform over time. These dynamic Gaussians are guided by a 4D Motion Scaffold, a sparse graph of trajectory nodes, enabling the interpolation of dense motion trajectories and features for each Gaussian efficiently. A separate set of static 3D Gaussians represents static background elements.

Feature4X introduces a unified latent feature embedding, distilled from various foundational 2D models, which facilitates multiple downstream tasks such as segmentation, scene editing, and visual question answering (VQA).

Specifically, Feature4X extracts video segment features from the InternVideo2-Chat model, a foundation model fine-tuned for video question answering.

This unified feature field is directly used by the InternVideo decoder for inference, bypassing the video encoding step entirely. This approach significantly improves inference efficiency and retains comprehensive structural information from the 4D scene representation, which surpasses the context available from the original 2D videos alone. Consequently, this method enhances downstream tasks by providing richer spatiotemporal context and semantic consistency.

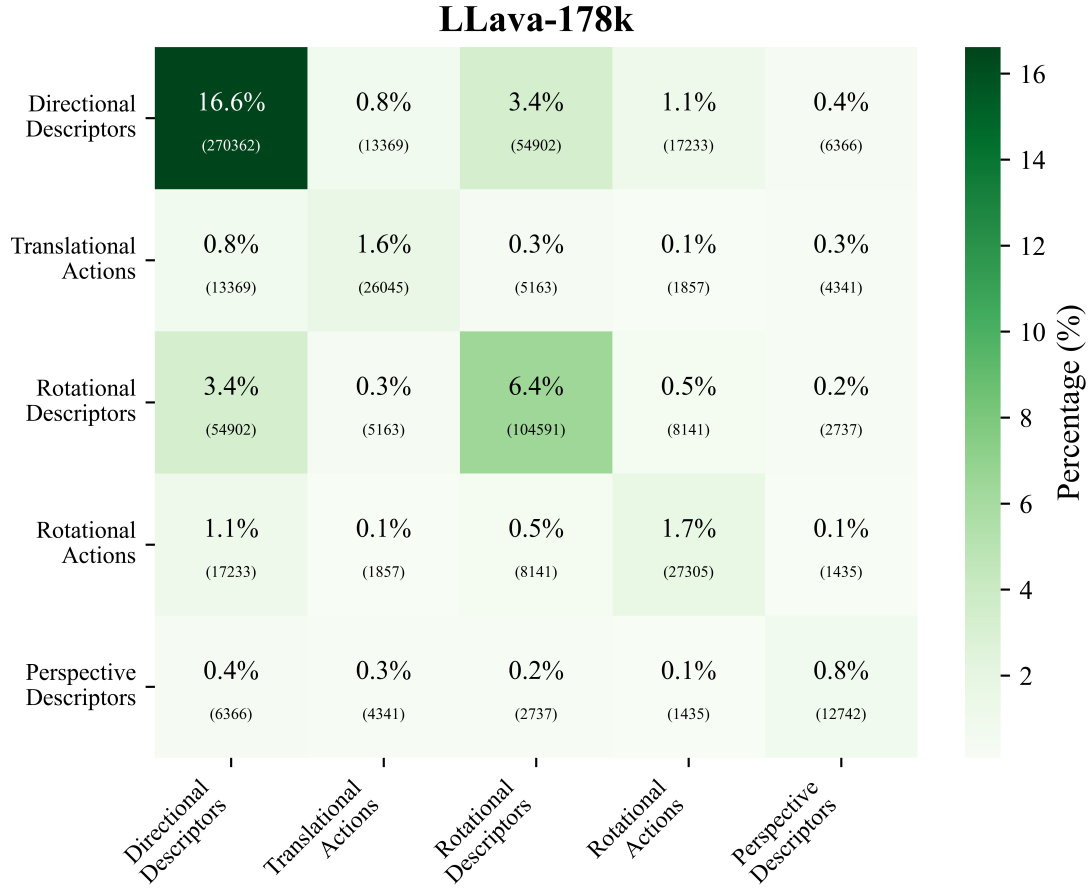


Figure B. Heatmap of Occurrences of Spatial-Temporal Terms in LLava-178k

(Note: LLava-178k actually comprises over 1.6 million samples, as we combine many of the available dataset splits within this collection.)

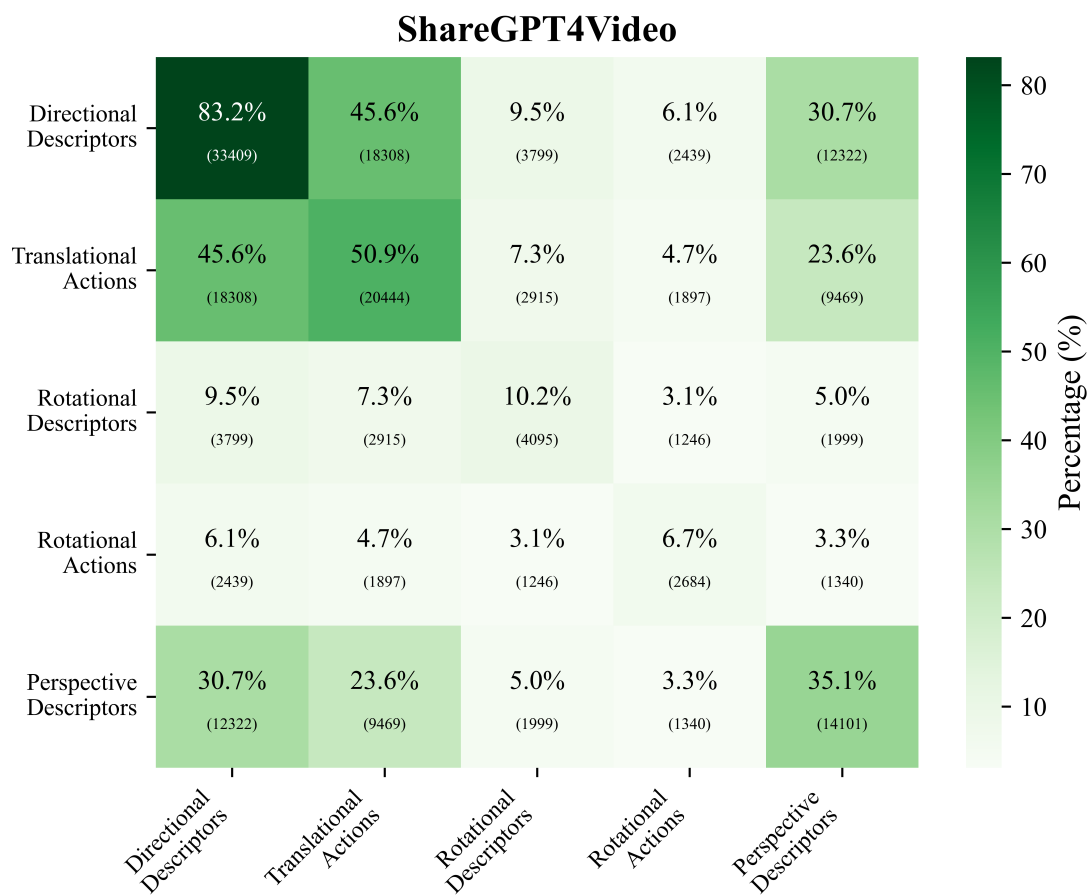


Figure C. Heatmap of Occurrences of Spatial-Temporal Terms in ShareGPT4Video

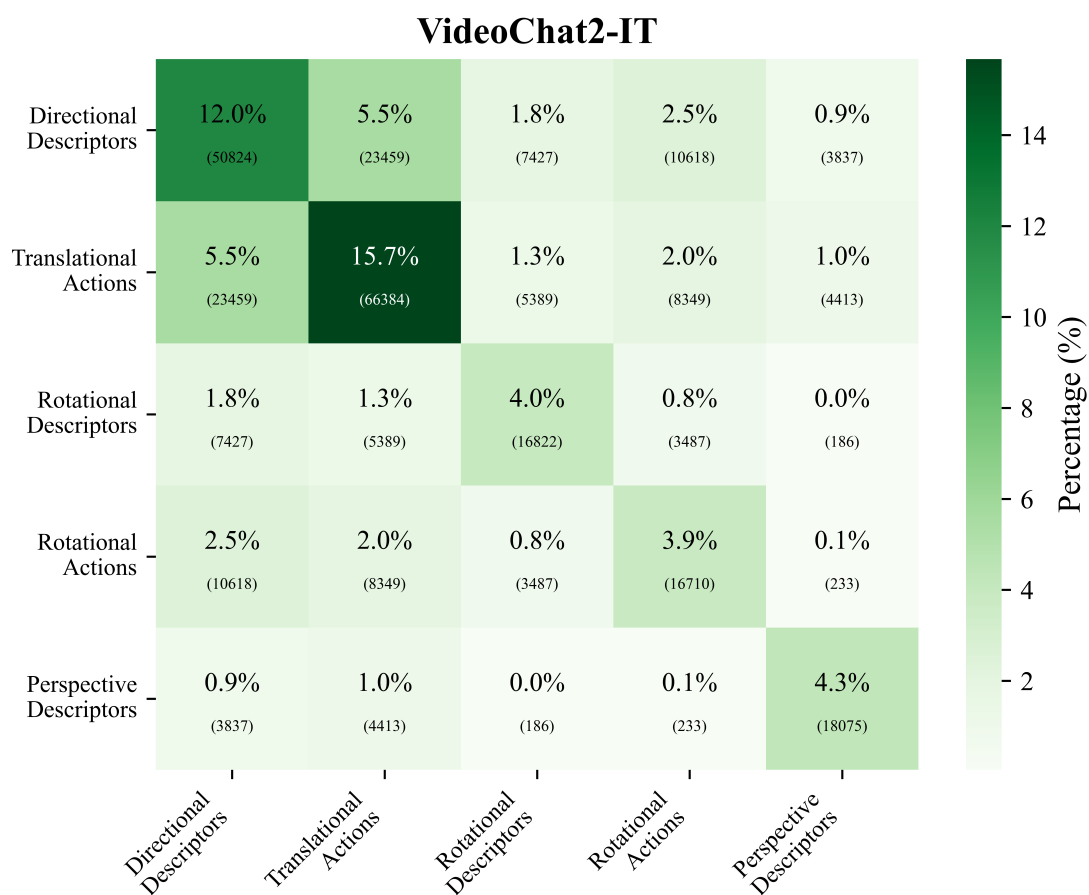


Figure D. Heatmap of Occurrences of Spatial-Temporal Terms in VideoChat-IT

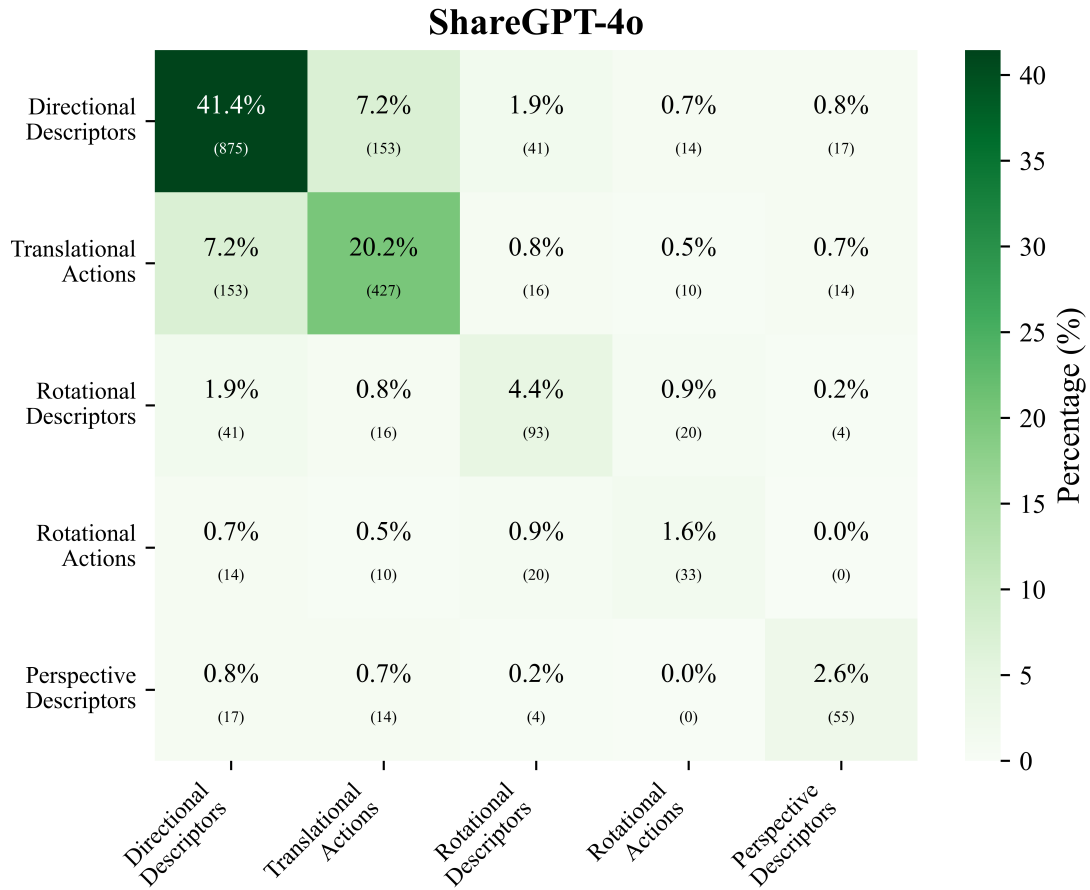


Figure E. Heatmap of Occurrences of Spatial-Temporal Terms in ShareGPT-4o

ShareGPT-4o



System Prompt:

Generate video descriptions that include details of visual effects, character actions, and movement of people/objects within frames. describe this video and its style to generate a description. pay attention to all objects in the video. do not describe each frame individually. instead of describing the imaginary content, only describing the content one can determine confidently. do not describe the contents by itemizing them in list form.

Caption:

A: the video captures a person **moving left**, near the kitchen table. the individual, wearing an apron, interacts with various items on the table, including a white mug, a cup with blue circular patterns, and a metallic container. throughout the video, the person uses their hands to manipulate the objects, **moving** them around the table. the table's surface and the items on it remain in focus while the background, which appears to be a kitchen setting, includes additional items like a water container that is **turned** and groceries. the video focuses on the hands and objects with no other significant movement, maintaining a steady and close-up **viewpoint** throughout the sequence.



- Directional Descriptors
- Translational Actions
- Rotational Descriptors
- Rotational Actions
- Perspective Descriptors

Figure F. Example of a clip with multiple target categories. Spatiotemporal grounding remains a challenge, as the direction descriptor is **incorrect**, and the translational actions provide limited insight, primarily indicating that objects/subjects are not static.

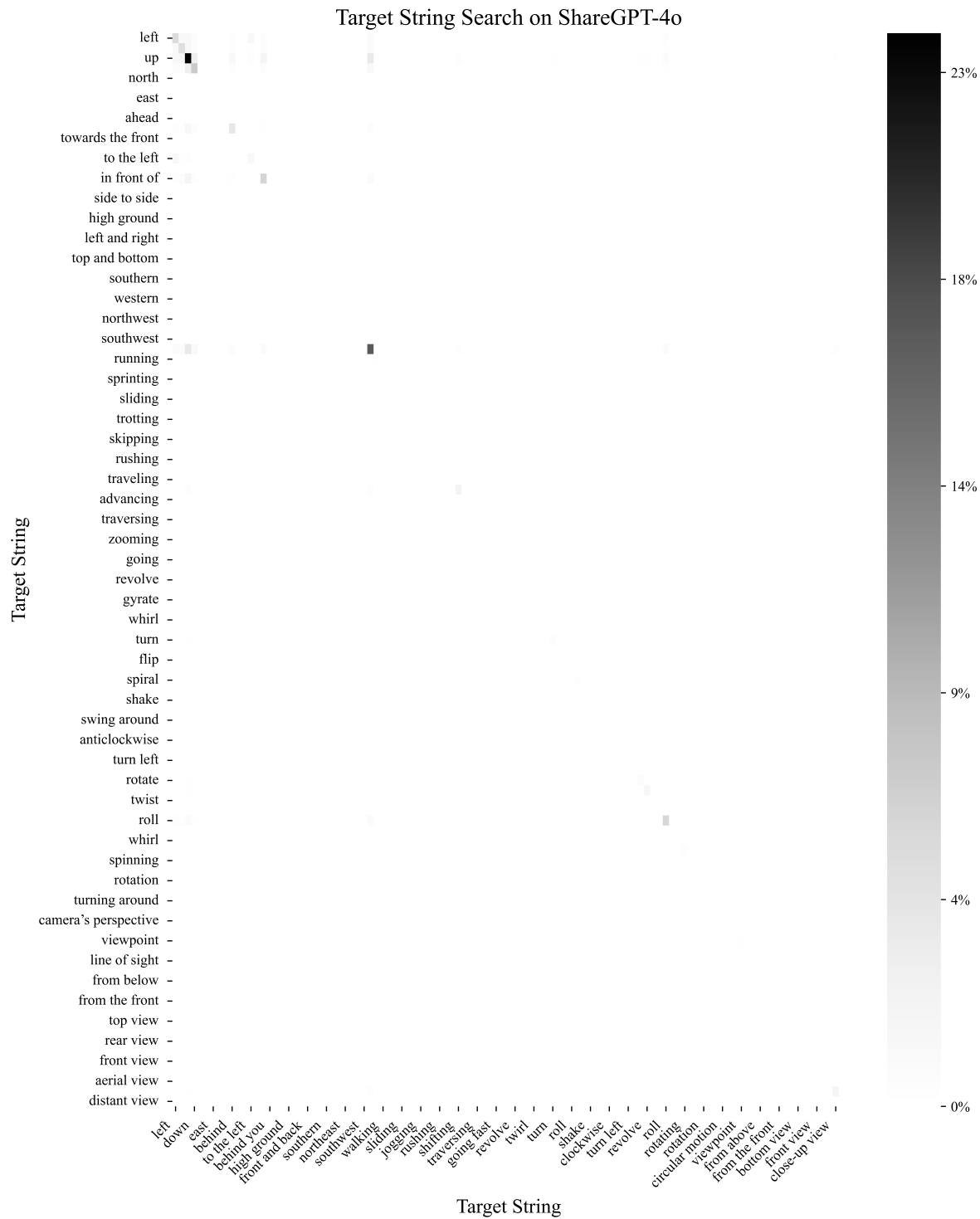


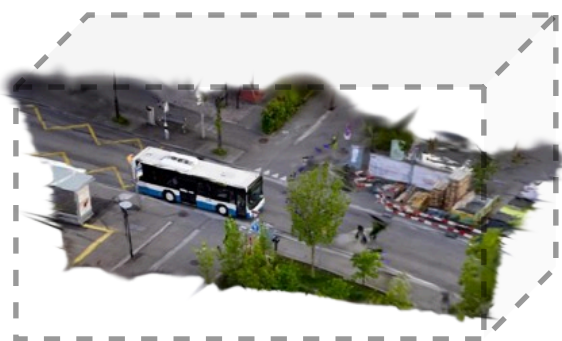
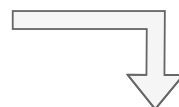
Figure G. Heatmap to visualize the prevalence of co-occurring target strings across the ShareGPT-4o dataset, informing our evaluation of spatiotemporal grounding in video instruction data.



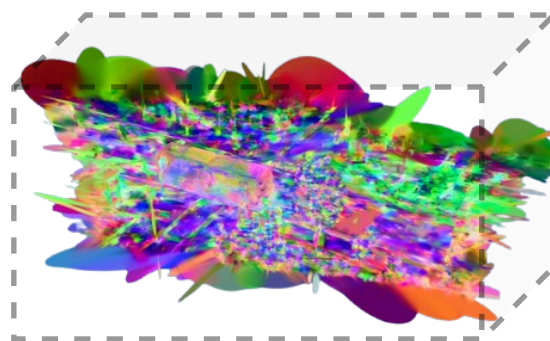
2D Video



Feature4X



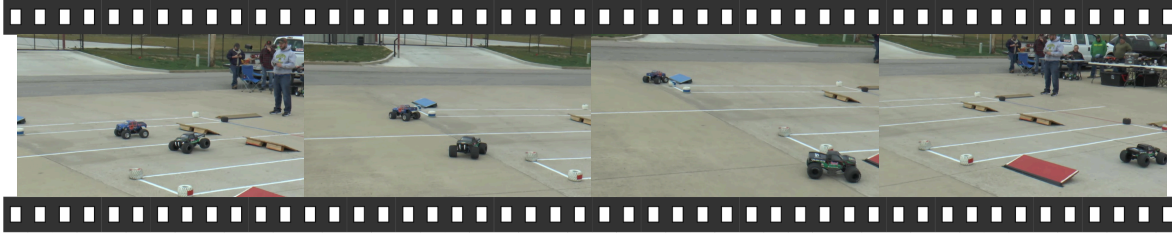
Global View Video



Lifted Internvideo2 Global Features

Figure H. General Input-Output Architecture of Feature4X

Is the black and green toy truck going clockwise or counter-clockwise around the course?
A. moving straight and not turning **B. counter-clockwise** C. clockwise D. not moving



VLM Responses	
GPT-4o	<p>Chain of Thought Response: To determine the direction the black and green toy truck is moving around the course, we need to analyze the sequence of images:\n\n1. Initial Position: In the first few images, the truck is on the right side of the course.\n\n2. Movement Observation: As the sequence progresses, the truck moves towards the lower part of the images, passing obstacles on its left side.\n\n3. Further Movement: The truck continues its path to the bottom left of the images, indicating it is following a path that curves to the left.\n\n4. Final Position: By the final images, the truck is seen moving off the course, having taken a wide left turn.\n\nBased on these observations, the truck is moving in a pattern that makes left turns, indicating it is following a counter-clockwise direction around the course.\n\nTherefore, the final answer is: B</p> <p>Direct Output Response: B: counter-clockwise</p>
Gemini-2.5-Pro	<p>Chain of Thought Response: * Step 1: Observe the truck's position at the start of its run. * At 00:02, we see a wide shot of the course. The black and green toy truck is positioned in the right-hand lane, pointing away from the camera, preparing to move "up" the course.\n\n* Step 2: Observe the truck's movement in the top section of the course. * At 00:00, the truck is shown jumping over a ramp in the top section of the course. It is moving from the right side of the screen towards the left side. This indicates that after traveling up the right lane, it is crossing over to the left lane at the top of the course.\n\n* Step 3: Observe the truck's position at the bottom of the course. * At 00:01, the truck is seen at the bottom of the course. It has just come down the left lane and is making a left turn to re-enter the right lane and begin another lap.\n\n* Step 4: Synthesize the path. * The truck travels up the right-hand lane, turns left at the top, travels down the left-hand lane, and turns left again at the bottom. This complete path forms a loop.\n\n* Step 5: Determine the direction of the loop. * A path that involves moving up on the right and down on the left, with left turns connecting the straightaways, is a counter-clockwise direction.\n\nTherefore, the final answer is: B</p> <p>Direct Output Response: B</p>
Claude-Sonnet-4	<p>Chain of Thought Response: I need to track the movement of the black and green toy truck throughout the sequence of images to determine if it's moving clockwise, counter-clockwise, straight, or not moving at all.\n\nLet me observe the black and green truck's position across the frames:\n\n1. In the early frames, I can see the black and green truck is positioned in the lower portion of the course.\n\n2. As the sequence progresses, the truck moves toward the right side of the course.\n\n3. Then it continues moving toward the upper portion of the course.\n\n4. Later in the sequence, it moves toward the left side.\n\n5. Finally, it appears to be moving back toward the lower portion, completing what looks like a circular pattern.\n\nWhen viewed from above, if I imagine the course as a clock face:\n\n- The truck starts at roughly the 6 o'clock position (bottom).\n\n- Moves to about 3 o'clock position (right).\n\n- Then to about 12 o'clock position (top).\n\n- Then toward 9 o'clock position (left).\n\n- And back toward 6 o'clock position (bottom).\n\nThis movement pattern from bottom → right → top → left → bottom represents a clockwise rotation around the course.\n\nTherefore, the final answer is: C</p> <p>Direct Output Response: C</p>
DeepSeek-VL2	<p>Chain of Thought Response: In the image provided, we can observe a black and green toy truck on a marked course. Its position relative to multiple white cones, ramps, and white boundary lines indicates movement direction. By analyzing spatial orientation and the location of the truck relative to obstacles, it appears that the black and green toy truck is making a leftward, counter-clockwise curve among the hurdles. Therefore, the final answer is: B.</p> <p>Direct Output Response: B</p>

Figure I. Complete CoT and DO Responses of VLMs Models 1-4

Grok-2-Vision	<p>Chain of Thought Response: To determine the direction the black and green toy truck is moving, let's analyze the sequence of images step by step:\n\n1. First Image: The black and green truck is positioned on the right side of the course, near the starting line.\n2. Second Image: The black and green truck has moved slightly forward.\n3. Third Image: The black and green truck is moving further forward, and the blue truck is still ahead.\n4. Fourth Image: The black and green truck is now clearly moving forward, passing the blue truck.\n5. Fifth Image: The black and green truck continues moving forward.\n6. Sixth Image: The black and green truck is further along the course.\n7. Seventh Image: The black and green truck is now on the left side of the course, heading towards the top.\n8. Eighth Image: The black and green truck is at the top-left corner, starting to turn down.\n9. Ninth Image: The black and green truck is moving down on the left side of the course.\n10. Tenth Image: The black and green truck is further down on the left side.\n11. Eleventh Image: The black and green truck is now at the bottom-left corner, starting to turn right.\n12. Twelfth Image: The black and green truck is moving right along the bottom of the course.\n13. Thirteenth Image: The black and green truck is moving further right.\n14. Fourteenth Image: The black and green truck is now at the bottom-right corner, starting to turn up.\n15. Fifteenth Image: The black and green truck is moving up on the right side.\n16. Sixteenth Image: The black and green truck is further up on the right side.\n17. Seventeenth Image: The black and green truck is nearing the top-right corner.\n\nFrom this sequence, we can see that the black and green truck is moving in a pattern that takes it from the right side to the left side, then down, then to the right, and then up again. This movement pattern indicates a clockwise direction.\n\nTherefore, the final answer is: C</p> <p>Direct Output Response: C</p>
Llama-4-Maverick-17B-128E-Instruct	<p>Chain of Thought Response: To determine the direction in which the black and green toy truck is moving around the course, we need to analyze the given images.\n\n1. Observation of Images: The images provided show a series of snapshots of toy trucks being operated in an outdoor setting, presumably a parking lot that has been converted into a racing or obstacle course. The images are taken from different angles and at different times, capturing the movement of the trucks.\n2. Identifying the Black and Green Truck: We need to identify the black and green toy truck across the different images to track its movement.\n3. Tracking Movement: By examining the sequence of images, we can track the position and movement of the black and green truck relative to other objects and the course layout.\n4. Analyzing Direction: \n - In Image 1, the truck is seen near the center.\n - In Image 3, it has moved further away from the center towards the top-left.\n - In Image 4, it continues to move away from the center, now towards the top.\n - In Image 5, it is seen at the bottom-right, indicating it has moved in a certain direction around the course.\n5. Determining Clockwise or Counter-Clockwise Movement: To ascertain if the truck is moving clockwise or counter-clockwise, we need to observe its position relative to fixed points or other trucks over the sequence of images.\n6. Analyzing the Sequence: \n - The images suggest that the trucks are moving around the course in a manner that can be inferred from their positions in successive images.\n - The specific path taken by the black and green truck indicates it is moving in a particular direction around the course.\n7. Conclusion: Based on the progression of the truck's position across the images (from center to top-left, then top, and eventually bottom-right), and considering the layout of the course, we can infer the direction of its movement.\n\nGiven the observed movement from the images, the black and green toy truck appears to be moving in a counter-clockwise direction around the course as it progresses from one part of the course to another.\n\nTherefore, the final answer is: B</p> <p>Direct Output Response: B</p>
Llama-4-Scout-17B-16E-Instruct	<p>Chain of Thought Response: To determine the direction of the black and green toy truck, let's analyze its position in the images provided.\n\n1. Initial Position and Movement: The truck starts on the right side of the course in several images, indicating its initial position.\n2. Observing Movement: In images 2, 7, and 8, the truck's position relative to other objects and its progression can be seen. \n3. Determining Direction: \n - In image 2, it is airborne, heading towards the left side of the frame.\n - In image 7, it appears on the right side of the frame, closer to the starting point or a specific landmark, suggesting it has moved.\n - In image 8, it is on the left side of the frame but seems to have moved further.\n4. Course Layout and Direction: The course has various obstacles and is laid out in a way that vehicles move from one end to the other, navigating through or over obstacles.\n5. Conclusion: Given that the truck moves from the right side towards the left side of the images and considering the course layout, we can infer its direction.\n\nTherefore, the final answer is: B</p> <p>Direct Output Response: C</p>

Figure J. Models 5-7

Phi-4-Multimodal-instruct	<p>Chain of Thought Response: To determine the direction in which the black and green toy truck is moving around the course, we need to observe its path in the video. The black and green toy truck is seen moving around the course, and it appears to be moving in a counter-clockwise direction. This is evident from the way it navigates the ramps and obstacles, moving from one side of the course to the other in a reverse direction compared to the blue and red toy truck. Therefore, the black and green toy truck is moving counter-clockwise around the course. Therefore, the final answer is: B.</p> <p>Direct Output Response: B</p>
Phi-3.5-Vision-instruct	<p>Chain of Thought Response: Looking at the video frames, we see a black and green toy truck positioned at the start of what appears to be an agility course. Initially, the truck is placed on the left side, facing towards the bottom right corner of the screen. As time progresses, the truck moves along the course, which includes turns and jumps. Throughout the course, the truck's orientation changes, indicating it is making turns. After completing one lap, the truck is positioned at the far right side, facing the left corner. Based on the direction the truck is moving from the initial to final position, we can deduce that the barriers and obstacles are arranged in a manner that requires the truck to move counter-clockwise. Thus, the final answer is: B. counter-clockwise. Therefore, the final answer is: B.</p> <p>Direct Output Response: C</p>
InternVL2.5-38B	<p>Chain of Thought Response: To determine the direction in which the black and green toy truck is moving, let's closely examine the images:\n\n- In Image 2, the black and green toy truck is positioned slightly to the left of the center of the frame.\n- In Image 3, the truck has moved to the right side of the image.\n\nThis indicates that the truck moved from left to right. Given that it starts from the left side and moves to the right side, it is going clockwise around the course.\n\nTherefore, the final answer is: C</p> <p>Direct Output Response: B</p>
InternVL2.5-8B	<p>Chain of Thought Response: To determine the direction the black and green toy truck is moving around the course, let's closely observe the positioning and trajectory of the toys in the images:\n\n1. Image 1: Both the blue and black and green toy trucks are on the course.\n2. Image 2: The black and green truck is on the lefthand side of the ramps, heading towards the obstacles.\n3. Image 3: The black and green toy truck is now positioned more centrally and appears to be near the middle of the course, likely moving forward.\n4. Image 4: The black and green toy truck has moved off the course, closer to the perimeter.\n\nBased on the sequence of images observed:\n- The truck transitions from the far left to the center, which suggests movement towards the rear of the course.\n- Given the arrangement and the constraint of the track, an advancing movement towards the back translates logically to a counter-clockwise progression since the truck is moving away from the immediate path towards outer obstacles.\n\nTherefore, the black and green toy truck is moving counter-clockwise around the course.\n\nTherefore, the final answer is: B</p> <p>Direct Output Response: A</p>
Pixtral-12b	<p>Chain of Thought Response: To determine the direction in which the toy truck is moving, we need to consider the standard convention for clockwise and counter-clockwise directions.\n\n1. Understanding Clockwise and Counter-Clockwise: - Clockwise direction means moving in a circular path in the same direction as the hands of a clock, starting from the top-right and moving to the right, then down, and so on.\n - Counter-clockwise direction means moving in a circular path opposite the direction of the clock's hands, starting from the top-left and moving to the left, then up, and so on.\n\n2. Analyzing the Options: - Option A: moving straight and not turning – This implies that the truck is not following a circular path, and thus it does not have a directional component of being clockwise or counter-clockwise.\n - Option B: counter-clockwise – This suggests the truck is moving in the counter-clockwise direction.\n - Option C: clockwise – This suggests the truck is moving in the clockwise direction.\n - Option D: not moving – This implies that the truck is stationary and not moving at all.\n\n3. Matching the Given Information: - The information provided does not specify a circular path or any turning, indicating no clear direction of motion in a circular path.\n - The truck is explicitly described as "moving straight and not turning," which aligns with Option A.\n\nTherefore, the final answer is: A</p> <p>Direct Output Response: B</p>

Figure K. Models 8-12

Aria-Chat	<p>Chain of Thought Response: From the provided images, we can observe the movement of the black and green toy truck around the course. In the first image, the truck is moving away from the ramps towards the center of the course. In the second image, it is driving towards the next set of obstacles. By the third image, the truck is turning towards the left, indicating a counter-clockwise direction. This pattern continues, and by examining the sequence of its movement around the course, it consistently follows a counter-clockwise path.\n\nTherefore, the final answer is: B: counter-clockwise.</p> <p>Direct Output Response: B: counter-clockwise</p>
Qwen2.5-VL-7B-Instruct	<p>Chain of Thought Response: To determine the direction of the black and green toy truck, let's analyze the image step by step:\n\n1. Identify the truck's position: The black and green truck is on the track, and we need to observe its orientation and the path it appears to be taking.\n\n2. Analyze the track layout: The track is rectilinear with designated paths for the trucks. There are other trucks visible on the track, but we focus on the black and green one.\n\n3. Determine the truck's movement: The truck appears to be moving forward, not along the straight sections but rather following the curved path.\n\n4. Direction of movement: The curve in the track suggests a direction that can be described as moving from one side of the track to the other, indicating a turn.\n\n5. Direction of movement: If we consider the overall direction the truck is taking, it is moving from the top-left corner of the image towards the bottom-right corner.\n\n6. Conclusion: This implies that the truck is moving in a counter-clockwise direction around the course.\n\nGiven the analysis, the black and green toy truck is moving counter-clockwise around the course.\n\nTherefore, the final answer is: B</p> <p>Direct Output Response: B</p>
Qwen2.5-VL-72B-Instruct-AWQ	<p>Chain of Thought Response: To determine the direction of the black and green toy truck, let's analyze its movement in the image:\n\n1. Identify the truck's position and direction: The black and green truck is seen near the bottom right of the image. It appears to be moving towards the bottom of the frame.\n\n2. Follow the truck's path: If we trace the likely path of the truck, it seems to be heading down and then will likely turn left towards the middle of the frame.\n\n3. Determine the overall direction: Given the truck's trajectory, it appears to be moving in a direction that is consistent with a counter-clockwise movement around the course. It is not moving straight, clockwise, or staying still.\n\nBased on this analysis, the truck is moving counter-clockwise around the course. Therefore, the final answer is: B</p> <p>Direct Output Response: B</p>
Qwen2-VL-7B-Instruct	<p>Chain of Thought Response: To determine whether the black and green toy truck is going clockwise or counter-clockwise around the course, let's analyze the positions of the trucks:\n\n1. Current positions: The current positions of the trucks are as follows:\n\n2. Black and green truck: The black and green truck is on the right side of the course.\n\n3. Orange truck: The orange truck is on the left side of the course.\n\n4. Directions: The trucks are currently separate from each other, moving in different directions.\n\n5. Inference: Since the image is not sequential, we cannot determine the movement between frames. However, we can infer the direction based on the current position.\n\n6. Options: Given the options:\n\nA: moving straight and not turning\nB: counter-clockwise\nC: clockwise\nD: not moving\n\nSince the black and green toy truck is on the right side of the course and there is no visual indication that it is turning towards the left, it is likely moving counter-clockwise around the course.\n\nTherefore, the final answer is: B</p> <p>Direct Output Response: B</p>
Qwen2-VL-72B-Instruct-AWQ	<p>Chain of Thought Response: In the given image, the black and green toy truck is moving clockwise around the course. It appears to be in the process of crossing one of the ramps on the course.\n\nTherefore, the final answer is: C</p> <p>Direct Output Response: B</p>
VideoLLaMA3-2B	<p>Chain of Thought Response: Answer: C</p> <p>Direct Output Response: C</p>
VideoLLaMA3-7B	<p>Chain of Thought Response: The black and green toy truck is moving in a counter-clockwise direction around the course. It starts at the bottom of the frame, moves to the right, then turns left and continues moving in a counter-clockwise direction. Therefore, the final answer is: B.</p> <p>Direct Output Response: B</p>

Figure L. Models 13-19

InternVideo2.5-Chat-8B	<p>Chain of Thought Response: Answer: (C) clockwise</p> <p>Direct Output Response: B: counter-clockwise</p>
InternVideo2-Chat-8B	<p>Chain of Thought Response: The toy truck is moving straight and not turning, so it is not going clockwise or counter-clockwise. Therefore, the final answer is: D.</p> <p>Direct Output Response: B: counter-clockwise</p>
LLaVA-onevision-qwen2-7b-ov-chat-hf	<p>Chain of Thought Response: Initially, the black and green toy truck with large black tires navigates around the course in a clockwise direction, moving back and forth between two blue ramps. In the final scene, it starts moving towards the edge of the parking lot, partially on the grass next to the concrete surface, continuing its movement in a clockwise direction before exiting the frame. Therefore, the final answer is: C</p> <p>Direct Output Response: C</p>
LLaVA-NeXT-Video-34B-hf	<p>Chain of Thought Response: I'm unable to determine the direction of motion for the toy truck in this image. The image does not provide enough information to conclude whether the truck is moving clockwise or counter-clockwise around the course. Without more information or seeing the movement of the truck in action, it is impossible to determine its direction of motion accurately.</p> <p>Direct Output Response: B</p>

Figure M. Models 20-23