

Can Language Models Understand mmWave Data? Benchmarking Large Language Models for mmWave Radar-Based Human Understanding

Supplementary Material

We provide supplementary material related to the main paper, arranged as follows:

1. Additional Dataset Specifications (Appendix A)
2. Additional Prompt Examples (Appendix B)
3. Additional Experimental Setup (Appendix C)
4. Additional VLM vs LLM Analysis (Appendix D)

A. Dataset Specifications

A.1. Action Categories

To build a unified benchmark suitable for question-answering, we first integrate three heterogeneous mmWave sensing datasets (mmBody, mRI, and MM-Fi), which originally contain 100, 12, and 27 action classes, respectively. However, many of these actions either overlap or are semantically too fine-grained for consistent reasoning in a QA setting. Therefore, we manually reorganize and redefine their labels to ensure that each action category is distinctive and unambiguous for the interpretation of questions. Through this consolidation, we derive 86 unified action categories (Table 4), systematically grouped into four motion domains: Upper-body (52), Lower-body (11), Torso (7), and Full-body (16).

A.2. Frame Statistics

We construct our dataset by integrating three public mmWave sensing datasets, mmBody, mRI, and MM-Fi, which collectively contain approximately 680K frames. Since raw radar recordings include many frames without meaningful movement, we first segment the sequences into intervals where human actions are actually performed, resulting in 488K motion-active frames. To ensure semantic consistency across heterogeneous sources, we then further refine the dataset by retaining only the segments that align with our predefined unified action taxonomy, removing instances with ambiguous or mismatched labels. Through this multi-stage filtering and consolidation process, we obtain a final curated set of 485K validated frames that are both action-centric and annotation-aligned, forming a reliable foundation for downstream question-answering tasks.

A.3. Task-Oriented QA Design

We propose five question-answering tasks, ActRec, TrajCheck, ActOrder, ActNum, and LimbFocus, each targeting a different aspect of language-based generalization in large language models. For every task, we generate multiple semantically equivalent question formulations (five variations

per question) to assess the robustness of linguistic understanding. Representative examples for each task are provided in Table 5. During dataset construction, one question variant is randomly selected for each sample to ensure diverse phrasing while maintaining a consistent semantic intent.

B. Prompt Examples

This section presents the zero-shot prompts used for the five task types, ActRec, TrajCheck, ActOrder, ActNum, and LimbFocus, illustrating how each query is formatted for the LLM. In addition, we describe the Chain-of-Thought (CoT) prompting strategy for rationale generation, using ActOrder as a representative example.

B.1. Zero-Shot prompting

ActRec: Action Recognition The ActRec prompt is designed for zero-shot action recognition, providing the LLM with an instruction that specifies the task and answer format, a sequence of radar point-cloud frames spanning a defined start-to-end segment that encapsulates a single action, and a natural-language question with multiple candidate actions. The model must select exactly one option that best matches the underlying human movement.

ActRec Zero-shot Prompt Example

► Referece Video:



► LLM Input:

```
<Instuction> You are an expert in mmWave radar point cloud analysis. Given a sequence of mmWave radar point cloud frames, identify the human action being performed. Each frame contains points with coordinates [x, y, z, range velocity, amplitude, energy power]. Select exactly one option (A, B, C, or D) that best describes the action. </Instuction>
```

```
<Point Cloud> frame_270:[-0.2, 3.0, -0.4, 0.0, -0.0, 115.12], [-0.2, 3.0, -0.2, 0.0, -0.07, 108.0], ... </Point Cloud>
```

```
<Question> Which action is represented by this point cloud?
```

```
A. Sway B. Basketball Dribble C. Squat D. One Arm Forward E. Hug
```

```
</Question>
```

► LLM Output:

```
D. One Arm Forward
```

Subcategory	Activity
Upper body	Arm Handshake, Arm Swing, Arms Freestyle Swim, Arms Circle Motion, Arms Flap, Badminton Swing, Badminton Under Serve, Basketball Dribble, Basketball Shot, Both Hands Forward, Both upper-limb lateral extension, Boxing Guard, Chest Expansion (Horizontal), Chest Expansion (Vertical), Cup Drinking Pose, Fist Punch, Hand Rotation, Hands Knead, Hands Mop Clean, Hands Prayer Pose, Hands Toss Pose (Volleyball), Hands to Head, Hug, One Arm Forward, One Arm Pull, One Arm T-Pose, One Hand Pour, One Hand Raise, One Hand Salt Toss, One Hand Slap, One Hand Stir, One Hand Sweep, One Hand to Head, One arm front raise and the other arm side raise, Overhead Press, Overhead Throw, Palm Stop, Pan Cooking Motion, Picking up things, Pour Water Motion, Pull Up, Scoop Tap Motion, Single hand forward raise, Single upper-limb lateral extension, Single-hand lateral raise, Throwing, Volleyball Spike, both arms Lift Overhead, one arm Lift Overhead, both arms Lift to Chest, Barbell Row, Bench Press
Lower body	Front Kick, Front Lunge, Inside Kick, Kicking, Knee Fold, Knee Up, One Leg Stand, Side Kick, Side lunge, Single Leg Jump, Wide Leg Jump
Torso	Bowing, Diagonal Stand, Head move, Lean Back, Sway, Torso Tilt, Twist
Full body	Arms Open Stand, Attention, Exaggerated March Step, Feet Moderate Wide Stance, Feet Shoulder Width, Feet Wide Stance, High Knee Run, Jogging, Jumping Up, Mark Time, T-Pose, Unilateral upper-lower limb extension, Walk, Walk Backward, Walk Forward, Squat

Table 4. Action taxonomy by category type in mmWave-QA.

TrajCheck: Spatial Movement Verification The TrajCheck prompt targets binary movement verification by providing the LLM with an instruction, a radar point-cloud segment, and a question asking whether the person’s global position changes across the scene.

ActOrder: Temporal Action Sequencing The ActOrder prompt focuses on temporal reasoning by providing the LLM with two different actions within a radar point-cloud segment and using one as an anchor reference. The question asks which action occurs before, after, or at the same time as the anchor, requiring the model to infer the correct temporal order.

TrajCheck Zero-shot Prompt Example

► Referece Video:



► LLM Input:

<Instuction> You are an expert in mmWave radar point cloud analysis. Given a sequence of mmWave radar point cloud frames, determine whether the person’s overall position in space has changed across the sequence. Each frame contains radar points with [x, y, z, doppler, intensity]. Focus on the general location of the entire point cluster that represents the person. If the central area of the cluster shifts noticeably from its previous position over time, it means the person has moved. Minor vibrations, swaying, or small posture changes without a clear shift in overall location should not be considered movement. Respond only with “yes” if the person changed position, or “no” if they remained in roughly the same place. </Instuction>

<Point Cloud> frame_006413: [[0.0, 2.78, 0.13, 0.0, 4.0], [0.0, 2.81, 0.18, 0.0, 7.0], ... </Point Cloud>

<Question> Did the person walk or change location, indicating a shift of the whole body’s center position?

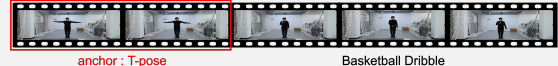
</Question>

► LLM Output:

Yes

ActOrder Zero-shot Prompt Example

► Referece Video:



► LLM Input:

<Instuction> You are an expert in mmWave radar point cloud analysis. Given a sequence of mmWave radar point cloud frames, identify the human action being performed. Each frame contains points with coordinates [x, y, z, range velocity, amplitude, energy power]. Select exactly one option (A, B, C, or D) that best describes the action. </Instuction>

<Point Cloud> frame_0:[[-0.2, 3.0, -0.4, 0.0, 0.07, 114.37],[[-0.2, 3.0, -0.2, 0.0, -0.0, 108.19], ... frame_120:[[-0.2, 3.0, -0.4, 0.0, 0.07, 111.19],[[-0.2, 3.0, -0.2, 0.0, 0.14, 112.87], ... </Point Cloud>

<Question> Which action happens just after T-Pose?

A. Volleyball Spike, B. Basketball Dribble, C. Hug, D. Leg Stand </Question>

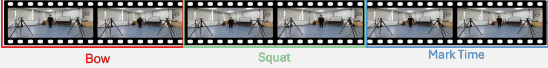
► LLM Output:

B. Basketball Dribble

ActNum: Action Cardinality Estimation The ActNum prompt evaluates the LLM’s ability to estimate action cardinality by providing a radar point-cloud segment that contains multiple actions and asking the model to determine how many distinct actions occur within the segment.

ActNum Zero-shot Prompt Example

► **Referece Video:**




► **LLM Input:**
 <Instuction> You are an expert in mmWave radar point cloud analysis. Given a sequence of mmWave radar point cloud frames, determine the requested count. Each frame contains points with coordinates [x, y, z, range velocity, amplitude, energy power]. Select exactly one option (A, B, C, or D). </Instuction>
 <Point Cloud> frame_000920: [[-0.4, 4.2, 0.0, 0.0, 0.07],[-0.4, 4.2, 0.2, 0.0, -0.0], ... frame_001062: [[-0.4, 4.2, 0.0, 0.0, -0.0],[-0.4, 4.4, -0.2, 0.0, 0.14], ... frame_001151: [[-0.4, 4.2, -0.4, 0.0, 0.07],[-0.4, 4.2, -0.2, 0.0, 0.07], ... </Point Cloud>
 <Question> How many distinct actions appear in this segment?
 A. 1, B. 2, C. 3, D. 4 </Question>

► **LLM Output:**
 C. 3

LimbFocus: Limb Motion Detection The LimbFocus prompt assesses localized motion understanding by asking the LLM to identify which specific limb or limb group is primarily active within the radar point-cloud segment.

LimbFocus Zero-shot Prompt Example

► **Referece Video:**



► **LLM Input:**
 <Instuction> You are an expert in mmWave radar point cloud analysis. Given a sequence of mmWave radar point cloud frames, identify the human action being performed. Each frame contains points with coordinates [x, y, z, range velocity, amplitude, energy power]. Select exactly one option (A, B, C, or D) that best describes the action. </Instuction>
 <Point Cloud> frame_270:[-0.2, 3.0, -0.4, 0.0, -0.0, 115.12], [-0.2, 3.0, -0.2, 0.0, -0.07, 108.0], ... </Point Cloud>
 <Question> Which single body part is primarily moving in this action?
 A. right arm, B. left arm, C. right leg, D. left leg, E. none </Question>


► **LLM Output:**
 B. left arm

B.2. CoT prompting

The CoT prompting setting extends the zero-shot prompt by adding an instruction after the question to elicit a brief rationale, guiding the LLM to analyze spatiotemporal evidence before outputting the reasoning and final answer in a predefined JSON format.

ActRec Prompt Example

► **Referece Video:**



► **LLM Input:**
 <Instuction> You are an expert in mmWave radar point cloud analysis. Given a sequence of mmWave radar point cloud frames, identify the human action being performed. Each frame contains points with coordinates [x, y, z, range velocity, amplitude, energy power]. Select exactly one option (A, B, C, or D) that best describes the action. </Instuction>
 <Point Cloud> frame_270:[-0.2, 3.0, -0.4, 0.0, -0.0, 115.12], [-0.2, 3.0, -0.2, 0.0, -0.07, 108.0], ... </Point Cloud>
 <Question> Which action is represented by this point cloud?
 A. Sway B. Basketball Dribble C. Squat D. One Arm Forward E. Hug
 </Question>

Given the radar point cloud data and the question, provide a brief reasoning for your answer. Analyze the spatiotemporal patterns in the point cloud to determine which action is being performed. Respond with ONLY a valid JSON object in this exact format (no markdown code blocks, no additional text):
 "Rationale": "Brief reasoning for the selected action", "Answer": "A. T-pose"

► **LLM Output:**
 "Rationale": "The point cloud data shows a consistent pattern of points moving forward and backward along the z-axis with minimal lateral movement, indicating a repetitive motion. The presence of points at varying heights suggests an arm movement rather than a full-body motion. The consistent velocity and amplitude changes are characteristic of a repetitive arm movement, such as extending one arm forward and then retracting it.", "Answer": "D. One Arm Forward"

C. Experimental Setup

The evaluation involves four commercial models: GPT-4o [36], GPT-5, Gemini 2.5-flash, and Gemini 2.5-pro [19]. These models are commonly used for multimodal reasoning tasks and provide a representative baseline for comparison. To ensure a fair and controlled comparison across modalities, we standardize the input configuration by uniformly

sampling 8 frames per sample for both the mmWave point-cloud sequences and the RGB videos. This prevents performance differences that may arise due to variations in temporal information. Model accuracy is computed by directly comparing each model’s predicted answer with the ground-truth label provided in the benchmark. All predictions are evaluated without any auxiliary post-processing or verification through external models (such as GPT-based answer checkers), ensuring that the reported accuracy reflects the models’ intrinsic reasoning capability.

D. Qualitative Analysis

In this section, we compare the generated rationales from video input and mmWave point-cloud input for each task, analyzing how their behaviors differ under six scene conditions: normal, furnished, rain, smoke, dark, and occlusion.

ActRec. Fig. 6 illustrates representative ActRec examples that compare how a VLM (RGB-based) and an LLM (mmWave point-cloud-based) interpret the same human actions across diverse scene conditions. Each row includes the reference RGB video, a multiple-choice question, and the rationales generated from both modalities, enabling a direct comparison of their reasoning. While the VLM can still infer actions reasonably well in smoke scenes by relying on coarse silhouette cues, its reasoning deteriorates significantly in darkness or occlusion, often producing incorrect predictions and hallucinated rationales that contradict the underlying motion. In contrast, the mmWave-based LLM maintains consistent and physically grounded interpretations across all conditions, highlighting the robustness of radar-guided reasoning when visual information becomes unreliable.

ActNum. In Fig. 7, we present representative ActNum examples that highlight the differences between a VLM with RGB inputs and an LLM with mmWave point-cloud inputs when estimating the number of distinct actions in a segment. Each row includes the reference RGB video, the corresponding question, and the rationales produced by both modalities. Unlike ActRec, the RGB-based VLM shows consistently poor performance across all conditions because it relies solely on visible motion transitions and thus becomes highly dependent on the number and spacing of input frames. When the silhouette appears static or motion cues are subtle, the VLM either fails to detect pose transitions or hallucinates nonexistent ones. In contrast, the mmWave-based LLM accurately distinguishes action changes by leveraging variations in point-cloud trajectories and Doppler signatures, which provide explicit cues about velocity and spatial movement even when visual differences are minimal. These examples highlight the inherent limita-

tions of RGB-based action counting and the robustness of radar-guided reasoning for ActNum.

ActOrder. Differences in how a VLM relying on RGB frames and an LLM using mmWave point-cloud data infer temporal action order are demonstrated through the ActOrder examples in Fig. 8. Each row contains the reference RGB video, the associated multiple-choice question, and the rationales generated by both modalities to allow a direct comparison of their temporal reasoning. Because the RGB-based VLM relies solely on silhouette shifts and visible limb trajectories, it often misinterprets the sequence of events when motion cues are subtle, partially occluded, or dominated by small upper-body movements. In several cases, this leads the VLM to incorrectly assume nonexistent transitions or to reverse the temporal order. In contrast, the radar-based LLM leverages stable Doppler patterns and point-cloud displacement trajectories, which provide explicit velocity and directional cues that directly encode temporal progression. As a result, the LLM consistently identifies the correct action ordering even under smoke, low-light, or ambiguous visual cues. These examples demonstrate the strength of radar-guided temporal reasoning and the susceptibility of RGB-based models to hallucinated or misaligned temporal interpretations.

LimbFocus. Representative LimbFocus examples are shown in Fig. 9, illustrating how a VLM using RGB inputs and an LLM using mmWave point-cloud inputs determine the primary body part involved in each action sequence. Each row includes the reference RGB video, a multiple-choice question, and the rationales generated from both modalities. Because the RGB-based VLM relies solely on visible limb silhouettes, it often fails under smoke, darkness, or low-motion conditions, where limb boundaries are blurred or indistinguishable. This leads the VLM to incorrectly assume “none” or to misattribute movement to the wrong limb. In contrast, the radar-based LLM leverages fine-grained point-cloud displacement and Doppler fluctuations, which provide explicit directional and velocity cues—even when visual appearance is degraded. As a result, the LLM consistently identifies the correct active limb across all conditions, demonstrating the robustness of radar-guided body-part reasoning compared to vision-only models.


Task	Subcategory	Question
ActRec	All	<p>Q1. What movement is shown in this radar data?</p> <p>Q2. Analyzing the point cloud and movement patterns, what motion is being captured?</p> <p>Q3. Based on the spatiotemporal distribution of radar points, what type of body movement or gesture can be inferred from this sequence?</p> <p>Q4. Looking at this radar point cloud data, what action is the person performing?</p> <p>Q5. Which action is represented by this point cloud?</p>
ActNum	All	<p>Q1. How many distinct actions appear in this segment?</p> <p>Q2. Count the unique poses shown in this sequence.</p> <p>Q3. What is the number of different actions in this window?</p> <p>Q4. How many unique action types are present here?</p> <p>Q5. What count of distinct poses occurs in this segment?</p>
ActOrder	Previous	<p>Q1. What action occurs immediately before <i>{anchor}</i>?</p> <p>Q2. Which pose precedes <i>{anchor}</i>?</p> <p>Q3. Right before <i>{anchor}</i>, what is the action?</p> <p>Q4. Identify the pose directly before <i>{anchor}</i>.</p> <p>Q5. Which action comes just prior to <i>{anchor}</i>?</p>
	Present	<p>Q1. What is the first action in this segment?</p> <p>Q2. Identify the initial pose in this sequence.</p> <p>Q3. What is the middle action in this segment?</p> <p>Q4. At the midpoint of this segment, what is the action?</p> <p>Q5. What is the last action in this segment?</p>
	Next	<p>Q1. What action occurs immediately after <i>{anchor}</i>?</p> <p>Q2. Which pose follows <i>{anchor}</i>?</p> <p>Q3. Right after <i>{anchor}</i>, what is the action?</p> <p>Q4. Identify the pose that comes directly after <i>{anchor}</i>.</p> <p>Q5. Which action happens just after <i>{anchor}</i>?</p>
LimbFocus	single	<p>Q1. Which single body part is primarily moving in this action?</p> <p>Q2. Identify the one body part that is active in this segment.</p> <p>Q3. Which body part is involved in this action?</p> <p>Q4. Select the body part that moves during this interval.</p> <p>Q5. What is the single moving body part in this pose?</p>
	Arm/Leg	<p>Q1. Are the arms or legs primarily involved in this action?</p> <p>Q2. Which body region is active: arms or legs?</p> <p>Q3. Does this action primarily use arms or legs?</p> <p>Q4. Identify whether arms or legs are moving in this segment.</p> <p>Q5. Which limb type is involved: arms or legs?</p>
	Multi	<p>Q1. Which combination of body parts is moving in this action?</p> <p>Q2. Select all body parts that are active in this segment.</p> <p>Q3. Identify the body parts involved in this action.</p> <p>Q4. Which body parts move during this interval?</p> <p>Q5. What combination of limbs is used in this pose?</p>
TrajCheck	All	<p>Q1. Did the person walk or change location, indicating a shift of the whole body's center position?</p> <p>Q2. Has the person moved from one place to another rather than staying in the same spot?</p> <p>Q3. Is there evidence of locomotion, such as walking, where the person's body center travels across coordinates?</p> <p>Q4. Did the person's body centroid translate noticeably, consistent with walking or relocation?</p> <p>Q5. Has the person shifted their overall position across frames, indicating actual movement through space?</p>

Table 5. Overview of the five QA task types in mmWave-QA with exemplar question phrasing for each semantic objective.

ActRec

<Reference Video: normal>



 **Question:** Analyzing the point cloud and movement patterns, what motion is being captured?
Option: **A. Sway**, B. One Arm Forward, C. Diagonal Stand, D. Chest Expansion (Vertical)

<VLM outputs - reasoning based on RGB video frames>


"Rationale": "The subject keeps both arms extended while shifting their body laterally from side to side. No arm-raising or chest-expanding movements appear in the sequence, ruling out the other options. The motion aligns with a balanced left-right sway."
"Answer": **"A. Sway"** ✓

<LLM outputs - reasoning based on mmWave radar point-cloud inputs>

"Rationale": "The point cloud data shows a consistent pattern of points moving slightly forward and backward along the z-axis, with minimal lateral or vertical displacement. This pattern is indicative of a swaying motion, where the body moves gently back and forth without significant changes in height or arm position."
"Answer": **"A. Sway"** ✓

<Reference Video: Smoke>



 **Question:** Looking at this radar point cloud data, what action is the person performing?
Option: **A. Badminton Swing**, B. Scoop Tap Motion, C. One Hand Pour, D. both arms Lift Overhead

<VLM outputs - reasoning based on RGB video frames>


"Rationale": "The subject performs a fast, diagonal arm motion across the body, resembling the typical swinging trajectory seen in racket sports. The repeated arm extension and follow-through are consistent with a striking action rather than a scooping or pouring motion. There is no indication of both arms lifting overhead."
"Answer": **"A. Badminton Swing"** ✓

<LLM outputs - reasoning based on mmWave radar point-cloud inputs>

"Rationale": "The point cloud data shows a consistent pattern of points moving in a linear trajectory with varying velocities and amplitudes, suggesting a repetitive motion typical of a badminton swing. The data does not show the characteristics of a scoop tap motion, one hand pour, or both arms lift overhead, which would involve different spatial distributions and velocities."
"Answer": **"A. Badminton Swing"** ✓

<Reference Video: Dark>



 **Question:** Analyzing the point cloud and movement patterns, what motion is being captured?
Option: A. Attention, **B. Overhead Throw**, C. Head Move, D. Front Kick

<VLM outputs - reasoning based on RGB video frames>


"Rationale": "The subject maintains an upright posture with minimal limb movement, and the silhouette stays mostly stable across frames. The only noticeable motion is a slight adjustment of the head and upper torso, consistent with someone directing their attention rather than performing a dynamic action. This still and focused stance aligns best with an attention posture."
"Answer": **"A. Attention"** ✗

<LLM outputs - reasoning based on mmWave radar point-cloud inputs>

"Rationale": "The point cloud data shows a consistent pattern of points moving vertically and horizontally, with some points extending further in the z-axis, indicating a motion that involves a significant range of movement. This pattern is consistent with an overhead throw, where the arm moves in an arc above the head, causing points to extend upwards and forwards."
"Answer": **"B. Overhead Throw"** ✓

<Reference Video: Occlusion>



 **Question:** Looking at this radar point cloud data, what action is the person performing?
Option: A. Pan Cooking Motion, B. Feet Moderate Wide Stance, C. Palm Stop, **D. Bow**

<VLM outputs - reasoning based on RGB video frames>

"Rationale": "The subject's lower-body silhouette stays wide and stable across the frames, with both legs positioned apart in a steady stance. There is no noticeable upper-body bending or arm movement that would indicate bowing, pan-cooking, or a palm-stop gesture. The overall posture is dominated by a balanced, moderately wide lower-body stance."
"Answer": **"B. Feet Moderate Wide Stance"** ✗

<LLM outputs - reasoning based on mmWave radar point-cloud inputs>


"Rationale": "The point cloud data shows a consistent pattern of points with minimal movement in the x and y coordinates, indicating a stationary position. The z coordinates show slight variations, suggesting a vertical motion. This pattern is consistent with a 'Bow' action, where the person bends forward and then returns to an upright position, causing changes in the vertical dimension while remaining relatively stationary in the horizontal plane."
"Answer": **"D. Bow"** ✓

Figure 6. **ActRec**: Examples comparing VLM (RGB-based) and LLM (radar-based) reasoning across diverse scene conditions, showing how each modality interprets the same action under normal, smoke, occlusion, and dark environments.

ActNum

<Reference Video: normal>



 **Question:** How many distinct actions appear in this segment?
Option: A. 1, B. 2, **C. 3**, D. 4

<VLM outputs - reasoning based on RGB video frames>

“Rationale”: “The subject transitions through four clearly distinguishable poses: an initial neutral stand, a turn-and-reach posture, a wide lateral arm extension, and finally a forward-bending action. Each posture changes both arm configuration and body orientation enough to be considered a separate action. These distinct shifts indicate that the segment contains four unique actions.”

“Answer”: **“D. 4”** ✗

<LLM outputs - reasoning based on mmWave radar point-cloud inputs>

“Rationale”: “The point cloud data across frames shows variations in the z-coordinate and intensity values, indicating changes in the vertical position and possibly the speed or reflectivity of objects. These variations suggest different actions or movements. By analyzing the changes in z-coordinates and intensity across frames, we can identify distinct actions. The data shows at least three distinct patterns of movement or change, indicating three distinct actions.”

“Answer”: **“C. 3”** ✓

<Reference Video: Smoke>



 **Question:** Count the unique poses shown in this sequence.
Option: A. 1, B. 2, **C. 3**, D. 4

<VLM outputs - reasoning based on RGB video frames>

“Rationale”: “Across the sequence, the person alternates between two visually distinct poses: arms stretched out to the sides and arms lowered closer to the body. These two postures repeat but no additional unique poses appear. Thus, the sequence contains two unique poses.”

“Answer”: **“B. 2”** ✗


<LLM outputs - reasoning based on mmWave radar point-cloud inputs>

“Rationale”: “The point cloud data across frames shows variations in the spatial distribution of points, particularly in the x, y, and z coordinates. These variations suggest different poses or configurations of the object or scene being captured. By analyzing the changes in the point cloud structure across frames, it is evident that there are three distinct configurations or poses present in the sequence.”

“Answer”: **“C. 3”** ✓

<Reference Video: Dark>



 **Question:** What is the number of different actions in this window?
Option: A. 1, **B. 2**, C. 3, D. 4

<VLM outputs - reasoning based on RGB video frames>

“Rationale”: “The silhouette remains nearly identical across all frames, with no clear limb movement, posture change, or body orientation shift visible despite the low-light conditions. The subject appears to stay in a single static stance without transitioning into additional poses. Therefore, only **one** distinct action can be identified in this segment.”

“Answer”: **“A. 1”** ✗


<LLM outputs - reasoning based on mmWave radar point-cloud inputs>

“Rationale”: “The point cloud data across frames shows variations in the z-coordinate and intensity values, indicating different actions. However, the changes are subtle and mostly consistent, suggesting a limited number of distinct actions. The data likely represents two different actions based on the observed patterns.”

“Answer”: **“B. 2”** ✓

<Reference Video: Occlusion>



 **Question:** How many unique action types are present here?
Option: A. 1, **B. 2**, C. 3, D. 4

<VLM outputs - reasoning based on RGB video frames>

“Rationale”: “The frames all show the same blurred striped pattern with no visible changes in motion or human activity. Since there is no variation across the sequence, only one consistent visual state is present. Thus, the number of unique action types is one.”

“Answer”: **“A. 1”** ✗

<LLM outputs - reasoning based on mmWave radar point-cloud inputs>

“Rationale”: “The point cloud data across frames shows variations in the z-coordinate and intensity values, indicating different actions. However, the changes are subtle and mostly consistent, suggesting a limited number of distinct actions. The data likely represents two different actions based on the observed patterns.”

“Answer”: **“B. 2”** ✓

Figure 7. **ActNum**: Examples comparing VLM (RGB-based) and LLM (radar-based) reasoning across diverse scene conditions, showing how each modality interprets the same action under normal, smoke, occlusion, and dark environments.

ActOrder

<Reference Video: Normal>



Question: What is the first action in this segment?

Option: A. Attention, **B. One Arm Forward**, C. Feet Shoulder Width, D. Pivot Turn

<VLM outputs - reasoning based on RGB video frames>

"Rationale": "At the beginning of the segment, the person raises one arm forward while the rest of the body remains in an upright stance. This motion is distinctly visible before any widening of the legs or body turns occur. Based on the temporal order, this indicates that "One Arm Forward" is the first identifiable action in the sequence."

"Answer": "B. One Arm Forward" ✓

<LLM outputs - reasoning based on mmWave radar point-cloud inputs>

"Rationale": "The point cloud sequence exhibits repeated horizontal displacement near the upper torso region, with dense point clusters shifting laterally frame by frame. These regions also show clear Doppler fluctuations consistent with fast oscillatory motion, which is commonly associated with arm-based gestures. In contrast, the lower-body regions remain spatially stable, reinforcing the conclusion that the primary movement arises from arm activity."

"Answer": "B. One Arm Forward" ✓

<Reference Video: Smoke>



Question: What action occurs immediately after Pull Up?

Option: A. Knee Up, B. Side Kick, C. Barbell Row, **D. Chest Expansion (Horizontal)**

<VLM outputs - reasoning based on RGB video frames>

"Rationale": "Following the upward movement of both arms consistent with a Pull Up motion, the next frames show a clear lateral extension of the arms outward at shoulder height. This type of symmetric horizontal displacement near the upper body region is characteristic of Chest Expansion (Horizontal), often observed as the arms move from a bent to a fully open posture. There is no significant leg involvement or vertical lifting afterward, confirming the transition to a horizontal upper-body gesture."

"Answer": "D. Chest Expansion (Horizontal)" ✓

<LLM outputs - reasoning based on mmWave radar point-cloud inputs>

"Rationale": "The point cloud representation would initially show a vertical upward movement of both arms close to the torso, forming the Pull Up motion. Immediately after, the points in the upper-body region shift laterally outward in both directions, forming a horizontal arc near shoulder height. This symmetric expansion pattern with stable leg positions and significant horizontal displacement in the upper-torso plane aligns with the characteristics of Chest Expansion (Horizontal)."

"Answer": "D. Chest Expansion (Horizontal)" ✓

<Reference Video: normal>



Question: Which action happens just after inside kick?

Option: **A. Overhead Throw**, B Squat, C Side Kick, D. Hands Toss Pose (Volleyball)

<VLM outputs - reasoning based on RGB video frames>

"Rationale": "After the Inside Kick, the point cloud shows immediate arm movement upward and outward from shoulder level. The symmetrical Doppler pattern suggests a coordinated motion consistent with a Hands Toss Pose, while no significant leg activity is detected."

"Answer": "D. Hands Toss Pose (Volleyball)" ✗

<LLM outputs - reasoning based on mmWave radar point-cloud inputs>

"Rationale": "After the Inside Kick, the point cloud displays a rising motion originating from the upper body, with strong Doppler activity tracing a forward and downward arc. This movement pattern matches the overhead arm trajectory typical of an Overhead Throw. No lateral symmetry or leg motion is observed, ruling out other candidate actions."

"Answer": "A. Overhead Throw" ✓

<Reference Video: normal>



Question: What action occurs immediately before Hug?

Option: A. Head move, **B. One Hand to head**, C. Boxing Guard, D. Sway

<VLM outputs - reasoning based on RGB video frames>

"Rationale": "The frames show only subtle shifts in the curved pattern with no identifiable limb outlines or any clear arm or leg motion. The minimal and localized changes suggest a slight upper-body or head-level movement rather than a full limb action. Since no distinct arm or leg displacement can be observed, the most reasonable interpretation is head movement."

"Answer": "A. Head move" ✗

<LLM outputs - reasoning based on mmWave radar point-cloud inputs>

"Rationale": "Right before the hug motion, the movement pattern shows a small and concentrated shift near the upper body rather than any wide or symmetric arm action. The motion is directed toward the head area, without the lateral spread or forward extension that would indicate other gestures. This localized action most closely corresponds to bringing one hand to the head."


"Answer": "B. One Hand to head" ✓

Figure 8. **ActOrder**: Examples comparing VLM (RGB-based) and LLM (radar-based) reasoning across diverse scene conditions, showing how each modality interprets the same action under normal, smoke, occlusion, and dark environments.

LimbFocus

<Reference Video: Normal>



 **Question:** Which single body part is primarily moving in this action?
Option: A. right arm, **B. left arm**, C. right leg, D. left leg, E. none

<VLM outputs - reasoning based on RGB video frames>

"Rationale": "Across the frames, the subject consistently lifts and swings the left arm outward while the rest of the body remains mostly stationary. The right arm and both legs show minimal displacement, indicating that they are not the primary source of motion. The repeated lateral extension of the left arm makes it the dominant moving body part in this sequence."

"Answer": "B. left arm" ✓


<LLM outputs - reasoning based on mmWave radar point-cloud inputs>

"Rationale": "The point cloud data shows concentrated motion on the subject's left side, where repeated fluctuations in x-y coordinates appear in the region corresponding to the left arm. Doppler variations indicate rapid directional changes that are characteristic of arm swings rather than leg movements, and no significant displacement is observed around the lower-body regions. These patterns collectively suggest that the left arm is the primary moving body part."

"Answer": "B. left arm" ✓

<Reference Video: Smoke>



 **Question:** Does this action primarily use arms or legs?
Option: **A. arms**, B. legs

<VLM outputs - reasoning based on RGB video frames>

"Rationale": "The person is repeatedly raising both arms above their head while keeping the lower body stable. No visible movement or displacement is observed in the legs throughout the sequence. The motion is entirely focused on the arms, indicating they are the primary body part used in this action."

"Answer": "A. arms" ✓


<LLM outputs - reasoning based on mmWave radar point-cloud inputs>

"Rationale": "The point cloud data showcases a frequent pattern of points clustering around coordinates with x values ranging between 0.6 and 3.3, and high doppler values (indicative of fast movement) in most frames. This is characteristic of upper body (arm) movement as the arms are generally closer to the radar and move faster than legs, especially when measured in activities that involve gesture or reaching motion."

"Answer": "A. arms" ✓

<Reference Video: Dark>



 **Question:** Select the body part that moves during this interval.
Option: **A. right arm**, B. left arm, C. right leg, D. left leg, E. none

<VLM outputs - reasoning based on RGB video frames>

"Rationale": "Across all frames, the subject's posture and position remain unchanged, with no visible movement in any limb. Even under low-light conditions, there is no detectable shift in arm or leg positions that would indicate active motion. Therefore, no body part appears to be moving during this interval."

"Answer": "E. none" ✗


<LLM outputs - reasoning based on mmWave radar point-cloud inputs>

"Rationale": "The point cloud data showcases a frequent pattern of points clustering in the region associated with right-arm motion, particularly along the x-axis at mid-range distances from the radar. These clusters exhibit strong Doppler activity across frames, indicating rapid and repeated movement localized to the upper body. No notable activity is observed in regions corresponding to the legs or the left side, supporting the interpretation that the right arm is the primary moving body part."

"Answer": "A. right arm" ✓

<Reference Video: Dark>



 **Question:** Identify the one body part that is active in this segment.
Option: **A. right arm**, B. left arm, C. right leg, D. left leg, E. none

<VLM outputs - reasoning based on RGB video frames>

"Rationale": "All frames in this segment show a static background with no visible human figure or discernible motion. There are no distinguishable body parts or movement cues present that could indicate activity. As such, no body part appears to be active in this segment."

"Answer": "E. none" ✗

<LLM outputs - reasoning based on mmWave radar point-cloud inputs>

"Rationale": "The point cloud data shows concentrated motion in the region typically corresponding to the left arm, with consistent horizontal displacement patterns in the upper-left area of the coordinate space. Doppler shifts occur frequently in that zone, indicating fast and isolated arm movements while the lower body and right-side regions remain static. This suggests that the left arm is the dominant moving part in this segment."

"Answer": "A. right arm" ✓

Figure 9. **LimbFocus**: Examples comparing VLM (RGB-based) and LLM (radar-based) reasoning across diverse scene conditions, showing how each modality interprets the same action under normal, smoke, occlusion, and dark environments.