SURE-BENCH: SENSOR UNDERSTANDING AND REASONING WITH EVIDENCE BENCHMARK FOR HUMAN ACTIVITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent human activity benchmarks increasingly incorporate heterogeneous sensors for multimodal understanding and reasoning. However, existing evaluations primarily verify the final task output or answer. Recent works indicate that correct answers do not necessarily mean that a model relies on the intended evidence, as VQA models can exploit shortcuts while still producing the expected answer. To address this issue, we introduce the Sensor Understanding and Reasoning with Evidence Benchmark (SURE-Bench), a multimodal and multi-view benchmark that evaluates both answer correctness and sensor-evidence correctness. SURE-Bench contains synchronized thermal, event camera, LiDAR, and radar observations across 61 indoor activity scenarios, with calibrated motion capture providing physical ground truth. We construct 687 sensor-exclusive atomic questions, each answerable from exactly one target modality, and compose them into 220 compound questions requiring specific sensor pairs. The required sensor evidence is therefore known by construction, allowing us to evaluate whether a model identifies the modalities required to support its answer. Experiments with recent VLMs show that answer correctness and sensor-evidence correctness do not consistently agree. Specifically, models may answer correctly without identifying the required sensor pair, or conversely, identify the required sensor pair without answering the question correctly. To the best of our knowledge, SURE-Bench is the first benchmark for indoor human activity to quantitatively evaluate answer correctness and sensor-evidence correctness separately, supporting research toward multimodal models whose answers are grounded in the required sensor evidence.

1 INTRODUCTION

Human activity understanding across a wide range of scenarios benefits from sensing systems that capture complementary information about people and their surroundings. For instance, rehabilitation may involve measuring step length and foot clearance, detecting rapid tremor, or monitoring surface-temperature changes during exercise. These measurements vary across sensing modalities. For instance, Thermal cameras capture heat patterns, LiDAR provides metric geometry, event cameras capture fine temporal changes, and mmWave radar provides range and motion information. As a result, different queries about the same activity can require evidence from different sensors, and certain queries require information from multiple modalities.

To address the problem, the existing line of research for multimodal human activity benchmarks can be broadly divided into two evaluation schemes: (1) Task-driven and (2) QA-driven. In task-driven evaluation, the model is given sensor observations and produces the output defined by a specific task. For example, action recognition predicts an action label, pose estimation predicts human joint locations, while higher-level tasks may generate an activity description, order observed actions, or predict the next action Ofli et al. (2013); Chen et al. (2015); Liu et al. (2020); Kong et al. (2019); Bocus et al. (2022); Yuan et al. (2025); Jiang et al. (2026). The predicted output is then compared with the corresponding task ground truth. More recently, in QA-driven evaluation, the task is instead expressed through a natural-language question, and the model produces an answer that is compared with the reference answer Shin et al. (2026).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

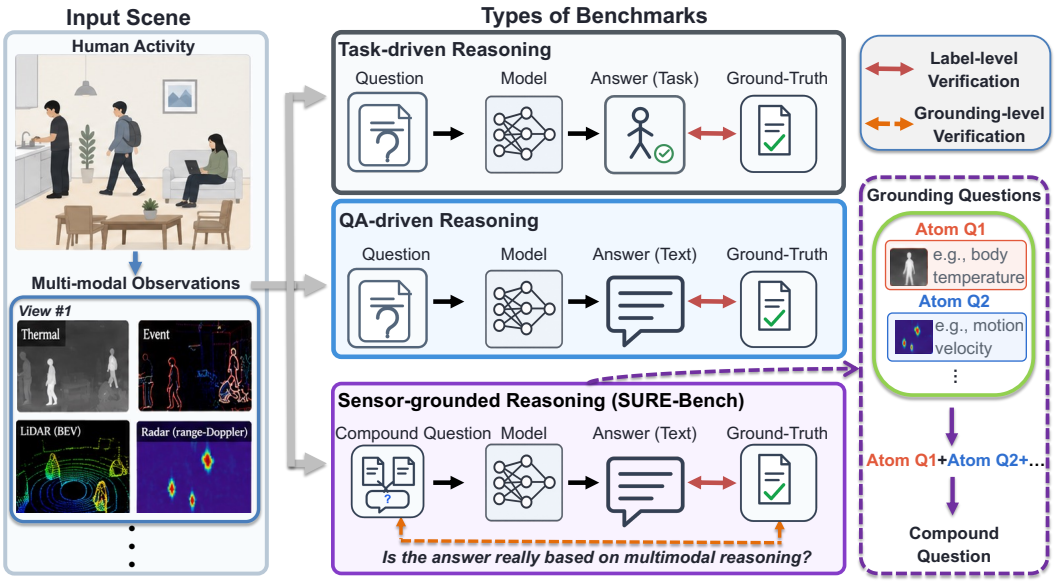

Figure 1: Existing human activity benchmarks primarily verify the final task output or answer. SURE-Bench additionally evaluates whether the model identifies the sensor evidence required to support its answer.

Despite these advances, both evaluation schemes primarily verify the final model output. Task-driven benchmarks determine whether the predicted label or state is correct, while QA-driven benchmarks determine whether the predicted answer is correct. However, a correct output alone may not reveal whether the model relies on the evidence intended by the task. In particular, existing VQA studies Dancette et al. (2021); Reich et al. (2023) show that models can exploit multimodal shortcuts or irrelevant visual evidence while still producing the expected answer. This issue becomes particularly important with heterogeneous sensors, where different modalities provide different physical information. For example, a model may answer a question about body temperature while identifying LiDAR rather than thermal observations as the supporting evidence. Therefore, evaluating only the final answer when several modalities are available does not necessarily reveal whether the model identifies the sensor evidence required to support it.

To address this issue, our core intuition is to separate answer correctness from sensor-evidence correctness. We introduce SURE-Bench, a multimodal and multi-view benchmark containing synchronized thermal, event camera, LiDAR point cloud, and radar observations across 61 indoor activity scenarios performed by 22 participants. The sensing system is spatially and temporally calibrated, with high-precision human pose ground truth recorded using motion capture devices, providing physical ground truth for verifying whether the model answer is correct. To provide sensor-evidence ground truth, we first construct sensor-exclusive atomic questions, where each question is only answerable from exactly one target modality. We then compose atoms from different modalities into compound questions related to human activities, where each question requires a specific sensor pair. Since the sensor requirement of each atom is known before composition, the sensor evidence required by the compound question is also known by construction. SURE-Bench can therefore evaluate whether a model reaches the correct answer and, separately, whether it identifies the sensor evidence required to support that answer. Figure 1 illustrates this distinction, and Table 1 compares SURE-Bench with representative multimodal human activity benchmarks.

SURE-Bench contains 907 expert-audited questions, including 687 sensor-exclusive atomic questions and 220 compound questions derived from the atomic set, which are used to evaluate recent VLMs. Our experimental results show that sensor misattribution ranges from 31.4 to 71.8 percent across the evaluated models, with correct attribution dropping below 10 percent for certain sensor pairs involving event cameras. These results expose a gap that final-answer evaluation is unable to capture. A model may answer a multimodal question while failing to identify the sensor evidence required to support that answer. Our main contributions are summarized as follows.

- We introduce SURE-Bench, a calibrated multimodal and multi-view benchmark for human activity reasoning from synchronized thermal, event camera, LiDAR, and radar observations, together with high-precision motion capture ground truth for answer verification.
- We introduce sensor-exclusive atomic questions and compound questions with human-expert audit. This enables explicit evaluation of whether a model identifies the sensor evidence required for its answer, which cannot be determined from answer correctness alone.
- We evaluate recent VLMs at both answer and sensor-evidence levels and reveal a substantial gap between answering multimodal questions and identifying the sensor evidence required to support those answers.

2 Related Work

Table 1: Comparison with representative human activity benchmarks. SURE-Bench explicitly defines the sensor evidence required by each question, allowing answer correctness and sensor-evidence correctness to be evaluated separately.

Benchmark	Venue	Heterogeneous Sensors	Multi-view	Physical / Pose GT	Language Reasoning	Explicit Sensor Requirement	Sensor-Evidence Verification
Task-driven benchmarks							
Berkeley MHAD Ofli et al. (2013)	WACV 2013	✓	✓	✓	–	–	–
UTD-MHAD Chen et al. (2015)	ICIP 2015	✓	–	–	–	–	–
NTU RGB+D 120 Liu et al. (2020)	TPAMI 2020	✓	✓	✓	–	–	–
MMAct Kong et al. (2019)	ICCV 2019	✓	✓	–	–	–	–
OPERAnet Bocus et al. (2022)	Sci. Data 2022	✓	✓	✓	–	–	–
OctoNet Yuan et al. (2025)	NeurIPS 2025	✓	✓	✓	–	–	–
CUHK-X Jiang et al. (2026)	MobiSys 2026	✓	–	✓	✓	–	–
QA-driven benchmarks							
mmWave-QA Shin et al. (2026)	CVPR 2026	–	–	–	✓	–	–
SURE-Bench		✓	✓	✓	✓	✓	✓

Task-driven human activity benchmarks. Multimodal human activity benchmarks have progressively expanded from visual and inertial sensing Ofli et al. (2013); Chen et al. (2015); Liu et al. (2020) to broader combinations of RGB, wearable inertial sensors, Wi-Fi, pressure sensing, and RF measurements Kong et al. (2019); Bocus et al. (2022). More recently, OctoNet Yuan et al. (2025) provides 12 heterogeneous modalities with multi-view sensing and motion-captured 3D pose labels. CUHK-X Jiang et al. (2026) further extends multimodal human activity evaluation from recognition to understanding and reasoning through tasks such as action description, temporal ordering, action selection, and next-action prediction. These benchmarks provide increasingly rich sensor observations and task ground truth. However, evaluation is primarily based on whether the model produces the expected task output.

QA-driven human activity understanding. Question answering provides a more flexible way to evaluate human activity understanding. Primarily based on RGB video, EgoVQA Fan (2019) introduced QA over egocentric videos, while EgoTaskQA Jia et al. (2022) evaluates descriptive, predictive, explanatory, and counterfactual reasoning about human tasks. Recent benchmarks further study human actions and interactions Wang et al. (2025b); Faure et al. (2026), spatio-temporal action relations Rodin et al. (2025), and reasoning about human goals and beliefs Li et al. (2025). More recently, mmWave-QA Shin et al. (2026) extends QA to non-RGB human sensing through mmWave measurements, but remains within a single sensing family. Multimodal QA has also been explored in autonomous driving using camera, LiDAR, and other scene information Qian et al. (2024); Tian et al. (2025); Sima et al. (2025); Mao et al. (2024); Marcu et al. (2024), showing the value of language-based evaluation over heterogeneous observations. However, QA-driven evaluation of human activities across multiple heterogeneous sensing modalities remains less explored.

Relation to SURE-Bench. SURE-Bench combines heterogeneous human activity sensing with QA-driven reasoning, but introduces an evidence-based evaluation target. Sensor-exclusive atomic questions define which modality provides the information each question requires, and their structured composition produces compound questions with known sensor requirements. Combined with calibrated multi-view observations and motion capture ground truth, this enables the separate evaluation of answer correctness and sensor-evidence correctness.

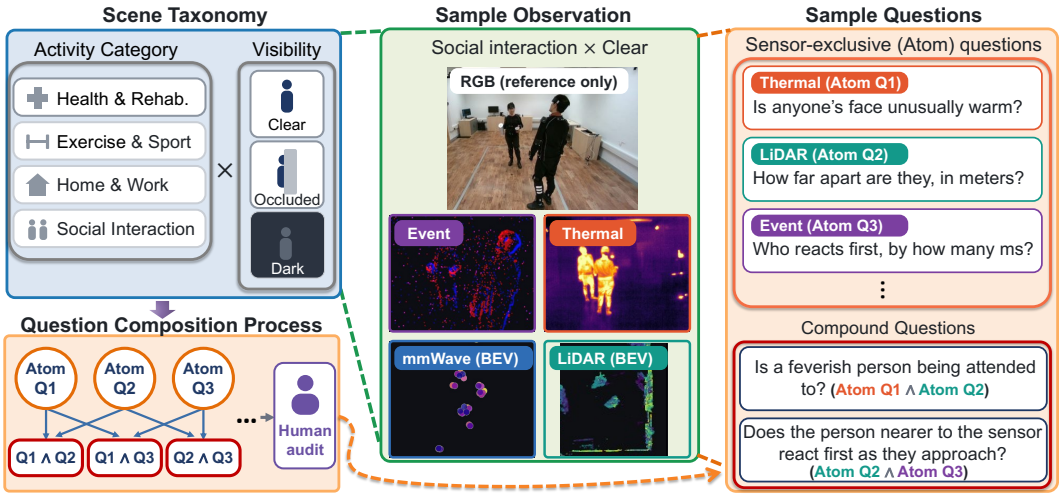

Figure 2: Overview of the SURE-Bench question construction. Indoor scenes cover different activity categories and visibility conditions and are recorded using synchronized heterogeneous sensors. We first construct sensor-exclusive atomic questions, where each question is answerable from one target modality. These atoms are then composed into compound questions that require a specific combination of sensors, followed by human verification. This construction makes the sensor evidence required by each compound question known by design.

3 SURE-BENCH

3.1 BENCHMARK OVERVIEW

SURE-Bench evaluates human activity reasoning from heterogeneous sensors at both the answer and sensor-evidence levels. We focus on four non-RGB target modalities: point cloud, event camera, thermal, and mmWave radar. As shown in Figure 2, the benchmark is organized by human activity and sensing condition. We consider four broad activity groups covering health and rehabilitation, exercise and sport, home and work, and social interaction. Activities are recorded under different visibility conditions, including clear, occluded, and dark scenes. The resulting synchronized observations capture the same activity through sensors that measure different physical properties. As a result, SURE-Bench contains 907 questions over 61 recorded indoor activities, including 687 sensor-exclusive atomic questions and 220 compound questions requiring information from two modalities. The following sections describe their construction and verification.

3.2 MULTIMODAL DATA COLLECTION

Data were collected in a $10\,\mathrm{m} \times 7\,\mathrm{m}$ indoor environment using four sensor rigs positioned around the capture area. As shown in Figure 3, the rigs provide overlapping views of the same activities using RGB-D, LiDAR, thermal, event camera, and mmWave radar sensors. Twelve ceiling-mounted Vicon cameras track the recording area and provide motion capture ground truth. Specifically, 16 anatomical joints per person, following Andriluka et al. (2014); Martinez et al. (2017) are recorded. All sensor streams are temporally aligned and spatially calibrated to the motion capture device's (Vicon) coordinate system. The calibrated observations and motion capture provide the physical reference used to determine whether a model answer is correct, independently of which sensor provides the evidence for that answer. Further details of the sensor specifications and synchronization are provided in the Appendix.

Figure 4 illustrates the necessity of the heterogeneous reasoning where common RGB perception would struggle, and why the modalities should be selectively chosen across sensing conditions. Under normal visibility (left), a person can be observed from all modalities. However, under low-light conditions, which could commonly occur in indoor scenarios, thermal and radar remain informative while the corresponding RGB and event observations become weaker.

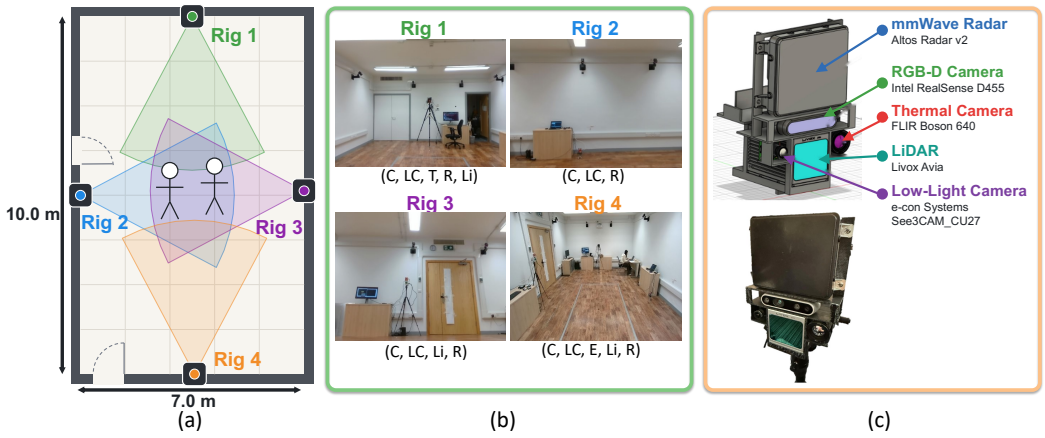

Figure 3: SURE-Bench data collection setup. (a) Floor map of the collection room. Four sensor rigs provide overlapping observations of the $7\,\mathrm{m} \times 10\,\mathrm{m}$ indoor capture area. (b) Example views from the four rigs and their available sensing modalities, where C, LC, E, T, R, and Li denote RGB-D, low-light RGB, event camera, thermal, mmWave radar, and LiDAR, respectively. (c) Example sensor hardware used for Rig 1. The multi-view setup provides complementary observations of the same human activities across heterogeneous sensors.

3.3 Sensor-Exclusive Atomic Questions

As Figure 2 shows, SURE-Bench's basic unit is a sensor-exclusive atomic question. Each atomic question is designed to require information from one target modality based on the unique information each modality provides. Specifically, point-cloud questions address metric distance, 3D geometry, and spatial relationships to the floor plane. Event-camera questions use fine temporal information, including high-frequency motion and short-latency event onset. Thermal questions cover surface temperature patterns, relative warmth, and residual heat after physical contact. Radar questions are based on range and Doppler measurements to capture human position and radial motion. These sensing capabilities allow different questions to be constructed based on the physical information provided by each modality.

Question Generation and Verification.

For each recorded scene, we provide the scene description and the physical information available from each sensing modality. To generate comprehensive questions, we utilized AI assistance to generate sensor-specific candidate questions. The LLM is used only to generate candidate questions and does not determine which questions are included in the benchmark.

Each candidate question is subsequently reviewed and refined by at least three domain experts with more than three years of experience and prior publications on the corresponding sensing modalities. More specifically, the reviewers examine the question against the provided activity context and sensor observations and verify three properties: (1) whether the question is supported by the recorded scene, (2) whether the information required to answer the question is available from the designated modality, and (3) whether the same information cannot be recovered from any of the other benchmark modalities. Questions that are ambiguous or can be answered from an unintended modality are revised or removed.

This process results in 687 sensor-exclusive atomic questions with verified answers and sensor requirements. As shown in Figure 2, these questions provide the building blocks for constructing compound questions that require information from multiple modalities.

3.4 Compound Question Construction

Compound questions are constructed by combining two verified sensor-exclusive atomic questions from the same activity scene, utilizing different target modalities. We formalize the required sensor evidence by defining $S(q)$ as the set of sensor modalities required to answer a question q. Given two

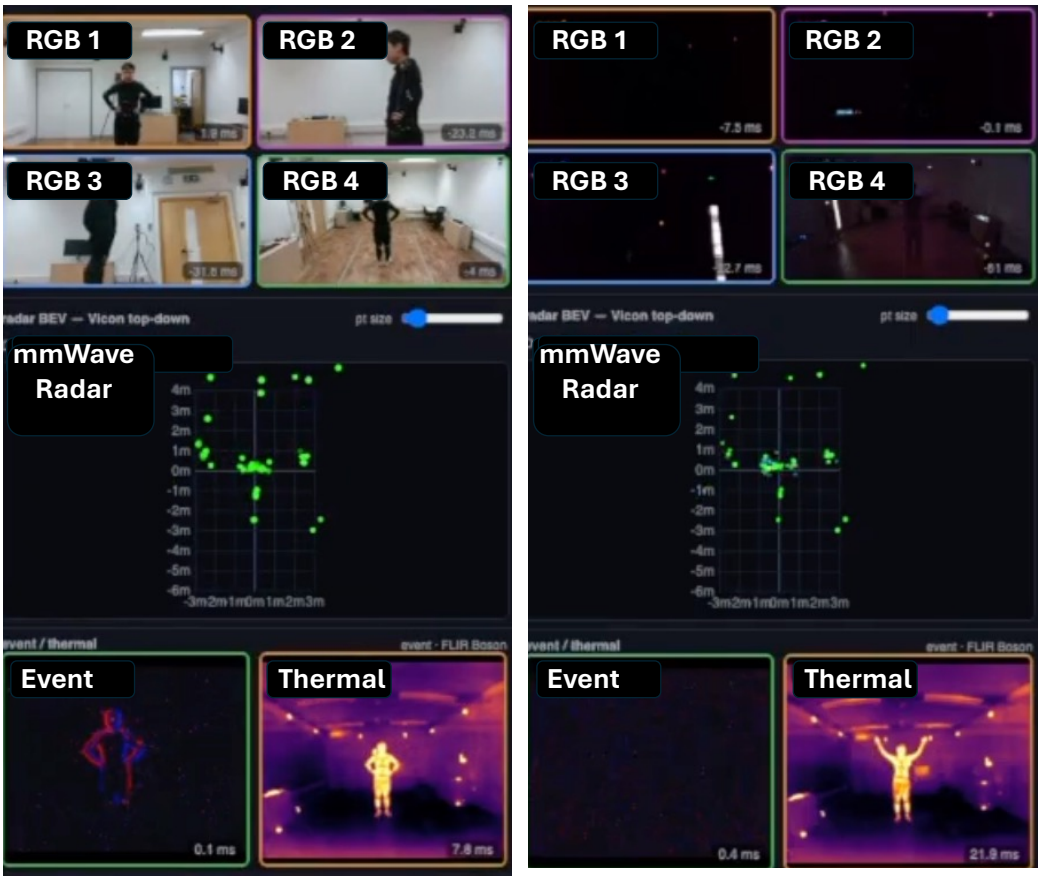

Figure 4: Examples of heterogeneous sensor observations for an Exercise & Sport activity across two SURE-Bench visibility categories. Left: Clear, where the person is observable across multiple modalities. Right: Dark, where thermal and radar remain informative while RGB and event observations become weaker.

atomic questions q_i and q_j requiring modalities s_i and s_j, respectively, the required sensor set for the resulting compound question q_{ij} is given by:

$$S(q_{ij}) = \{s_i, s_j\}, \qquad s_i \neq s_j. \tag{1}$$

The two verified atomic questions define the information that must be preserved in the compound question. We use LLMs to combine them into a single question that is natural and relevant to the corresponding activity scene while requiring information from both atoms. The LLM is used only to assist question formulation, while the required sensor pair is inherited from the verified atomic questions. Similar to the atomic question generation process, each generated compound question is subsequently reviewed by the domain experts. The reviewers verify that the question and the answer are directly grounded in both constituent atoms, are practical for the recorded activity scene, and still require information from both designated modalities. Questions that do not preserve these properties are revised or removed. Since the sensor requirement of each atomic question is verified before composition and preserved through expert review, the required sensor pair of the final compound question is known by construction.

The reference answer is derived from the verified constituent atoms and the calibrated 3D motion capture ground truth, while the required sensor pair is inherited from the constituent atoms. Beyond expert verification, we perform an adversarial sensor-exclusivity audit using an independent VLM. Specifically, for each atomic question, we remove the designated modality and provide the VLM with the remaining sensor observations. The VLM is then asked to answer the original question

Table 2: Distribution of atomic questions by answer type and compound questions by required sensor pair.

	Binary	Multiple-choice	Quantitative	Total
Point cloud	15	18	179	212
Event	100	51	50	201
Thermal	175	39	0	214
Radar	22	22	16	60
Atomic total	312	130	245	687

Compound pair	PC+T	E+PC	E+T	E+R	PC+R	R+T	Total
Questions	57	53	53	17	20	20	220

without being told which modality was removed. If the answer can be recovered from the non-target modalities, the question is considered to contain an unintended cue and is revised or removed. For compound questions, we similarly remove each required modality in turn and verify that the information inherited from both constituent atoms remains necessary after composition. This process results in 220 compound questions requiring different pairs of sensor modalities.

4 Experiments

4.1 Experimental Setup

Question and Answer Format. SURE-Bench contains three answer formats: binary, multiple-choice, and quantitative. Following the existing VQA evaluation practices Goyal et al. (2017); Zhu et al. (2016); Zhang et al. (2025); Yang et al. (2026), binary questions require a yes/no answer, multiple-choice questions select from the candidate answers specified in the question, and quantitative questions require a numerical answer evaluated within a predefined tolerance. Table 2 summarizes the question distribution. The 687 atomic questions contain all three answer formats, while the 220 compound questions use binary answers and are distributed across six required sensor pairs. PC, E, T, and R denote point cloud, event camera, thermal, and radar, respectively.

Models and Evaluation Protocol. We evaluate pretrained VLMs on SURE-Bench in a zero-shot setting. For atomic questions, each model receives the observation from the designated target modality. For compound questions, all four target modalities are provided. Specifically, visual observations are provided as separate labeled images and interleaved with the corresponding textual sensor representations, following the established multi-image VLM setting where multiple images and text are jointly processed within the same input Jiang et al. (2024). We utilize five publicly available VLMs that natively support multi-image and text inputs: Qwen2.5-VL Bai et al. (2025b), Qwen3-VL-32B Bai et al. (2025a), Gemma-4-31B Gemma Team (2026), InternVL3.5-38B Wang et al. (2025a), and Mistral-Small-3.2-24B Mistral AI (2025). The same questions, sensor representations, and prompting protocol are used across all models.

For compound questions, the required sensor pair is not provided to the model. Instead, the model receives all four modalities and is asked to answer the original question and identify the modalities required to support its answer. Providing all modalities prevents the input itself from revealing the required sensor pair and allows answer correctness and sensor-evidence correctness to be evaluated separately.

Sensor Representation. We represent each modality in a form that preserves the sensor information required by its questions. Thermal observations are provided directly as images, following recent VLM evaluation on thermal imagery Shrivastava et al. (2026). Point clouds are represented using a metric bird's-eye-view following prior vision-language work using LiDAR BEV Cao et al. (2024), together with sampled XYZ coordinates to preserve metric 3D information. Following mmWave-QA Shin et al. (2026), radar measurements are represented as structured numerical text containing spatial coordinates and Doppler velocity. Event-camera observations are represented using temporally ordered event visualizations, where each visualization accumulates events over a fixed temporal

Table 3: Overall zero-shot performance on SURE-Bench.

Model	Atomic Answer ↑ (687)	Compound Answer ↑ (220)	Evidence Acc. ↑ (220)	Joint Acc. ↑ (220)	Evidence|Correct ↑
Qwen2.5-VL	39.9	7.3	45.0	5.5	75.0
Qwen3-VL-32B	48.9	16.4	52.3	10.0	61.1
Gemma-4	53.7	16.8	68.6	13.2	78.4
InternVL3.5	40.2	12.7	53.2	8.2	64.3
Mistral 3.2	38.4	8.2	28.2	2.3	27.8

Table 4: Sensor-evidence accuracy by required sensor pair.

Model	PC+T (57)	E+PC (53)	E+T (53)	E+R (17)	PC+R (20)	R+T (20)
Qwen2.5-VL	61.4	47.2	22.6	52.9	55.0	40.0
Qwen3-VL-32B	63.2	66.0	39.6	23.5	60.0	40.0
Gemma-4	66.7	67.9	84.9	52.9	50.0	65.0
InternVL3.5	71.9	28.3	64.2	47.1	40.0	55.0
Mistral 3.2	73.7	1.9	5.7	23.5	20.0	45.0

window, consistent with prior event-language work that models the spatio-temporal structure of event streams Liu et al. (2025).

For compound questions, we use a fixed output format with separate answer and evidence fields. The answer field contains the response to the original question, while the evidence field contains the modalities that the model identifies as necessary to support that answer. We ask for the required supporting modalities rather than which modalities the model internally used, since the former can be directly compared with the sensor pair defined during benchmark construction.

4.2 Evaluation Metrics

Following our motivation in Sec. 1, we evaluate answer correctness and sensor-evidence correctness separately. For atomic questions, binary and multiple-choice answers are evaluated by exact correctness, while quantitative answers are considered correct when the predicted value falls within the predefined tolerance. For compound questions, answer correctness measures whether the model produces the correct final answer to the compound question. Sensor-evidence correctness is evaluated independently using exact set match. A prediction is correct only when the model identifies both required modalities without selecting an additional modality. We additionally report joint accuracy, which requires both the compound answer and sensor evidence to be correct. Evidence|Correct measures sensor-evidence accuracy among correctly answered compound questions. Unlike the overall evidence accuracy, this metric considers only the subset with a correct compound answer and measures how often the required sensor pair is also identified correctly.

Detailed atomic results by target modality and answer format, including binary and quantitative performance, are reported in Appendix A.1.

4.3 Modality-Evidence Evaluation

Our motivation of this experiment is to examine whether answer correctness and sensor-evidence correctness agree on the same compound questions. Table 3 reports the overall answer and evidence results, while Table 4 reports evidence accuracy by required sensor pair. PC, E, T, and R denote point cloud, event camera, thermal, and radar, respectively.

As shown in Table 3, answer accuracy decreases from 38.4–53.7% on atomic questions to 7.3–16.8% on compound questions. Sensor-evidence accuracy ranges from 28.2% to 68.6% and does not always correspond to answer correctness. Among correctly answered compound questions, the required sensor pair is identified correctly in 27.8–78.4% of cases. Conversely, evidence accuracy is substantially higher than joint accuracy for all models, showing that the required sensor pair can also be identified when the final answer is incorrect. Thus, answer correctness and sensor-evidence correctness can disagree in either direction. Table 4 further reports sensor-evidence accuracy by the required sensor pair. PC+T is relatively consistent across models, ranging from 61.4% to 73.7%. Other pairs vary considerably across models. For example, E+PC ranges from 1.9% to 67.9%, and E+T from 5.7% to 84.9%. This suggests that sensor attribution depends not only on the individual modalities but also on their combination.

Table 5: Modality necessity on atomic questions. Target uses the designated modality, while Non-target uses the remaining three. The gap indicates the difference in accuracy.

Model	Target ↑	Non-target	Gap ↑
Qwen2.5-VL	39.9	37.1	+2.8
Qwen3-VL-32B	48.9	43.4	+5.5
Gemma-4	53.7	40.3	+13.4
InternVL3.5	40.2	39.3	+0.9
Mistral 3.2	38.4	40.2	−1.8

Table 6: Modality necessity on compound questions under different combinations of required and non-required modalities.

Model	All	Required Only	Drop Required	Drop Irrelevant	Non-required Only
Qwen2.5-VL	7.3	10.5	6.4	8.6	6.4
Qwen3-VL-32B	16.4	17.3	15.0	16.8	10.9
Gemma-4	16.8	21.8	14.1	18.6	10.5
InternVL3.5	12.7	10.9	11.4	12.3	9.1
Mistral 3.2	8.2	7.7	8.6	9.1	7.7

4.4 Modality-Necessity Analysis

As shown in Sec. 4.3, identifying the modalities required to answer a question does not establish whether those modalities contribute to the answer. To investigate this further, we perform modality ablations to examine how answer accuracy changes when the available sensor evidence is altered.

For atomic questions, *Target* provides only the designated modality, while *Non-target* excludes the designated modality and provides the other three modalities. The gap between the two conditions measures how answer accuracy changes when the designated sensor is unavailable. For compound questions, *All* provides all four modalities, while *Required Only* provides the required pair. *Drop Required* and *Drop Irrelevant* remove one modality inside and outside the required pair, respectively, with results averaged over the two possible removals. *Non-required Only* provides the two modalities outside the required pair.

Table 5 shows mixed effects for atomic questions. Removing the target modality reduces accuracy by 0.9–13.4 percentage points for four models, but increases accuracy by 1.8 points for Mistral 3.2. The compound ablation gives a more consistent result. Across all five models, removing a required modality leads to lower accuracy than removing an irrelevant modality. The difference is largest for Gemma-4 (14.1% vs. 18.6%). Qwen2.5-VL, Qwen3-VL-32B, and Gemma-4 also perform better with only the required pair than with all four modalities. Together, these results show that required sensor information affects compound answering, while providing additional modalities does not necessarily improve performance.

5 Conclusion

In this work, we introduced SURE-Bench, a multimodal and multi-view benchmark for human activity reasoning from heterogeneous sensors. SURE-Bench constructs compound questions from verified sensor-exclusive atomic questions, providing known sensor requirements in addition to reference answers. This enables answer correctness and sensor-evidence correctness to be evaluated separately. Our experiments with recent VLMs show that identifying the modalities required to answer a question does not always correspond to answer correctness. The modality ablations further show that removing required sensor information generally affects compound answering more than removing irrelevant information. SURE-Bench provides a framework for evaluating these aspects separately and supports research toward multimodal models whose answers are grounded in the required sensor evidence.

AI use statement

In this work, we used generative AI tools to (1) retrieve BibTeX entries, (2) draft parts of the paper, (3) generate a part of a figure, and (4) generate the initial candidate set of evaluation questions. We have not used generative AI tools for other tasks indicated in the ICLR 2027 AI Policy for Authors. Specifically, we provided the AI with several titles of relevant works we were aware of and asked it to find the corresponding BibTeX entries for citations. After retrieval, we double-checked that all retrieved BibTeX entries were correct. For drafting, we used the AI tool to draft the related work section first, with instructions for subsections. After drafting, the authors heavily refined it to fit the

paper's context. For the figure, we used the AI to generate a conceptual drawing of the scene. More specifically, the conceptual drawings produced with the AI's assistance are in the left box named "Input Scene" in Figure 1. We did not use the AI for any other figures, including other parts of Figure 1. The process of utilizing AI for the initial candidate question set, followed by substantial audit from the human experts, is extensively described in Sec 3.3 and Sec 3.4.

REFERENCES

Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3686–3693, June 2014. doi: 10.1109/CVPR.2014.471.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Mohammud J. Bocus, Wenda Li, Shelly Vishwakarma, Roget Kou, Chong Tang, Karl Woodbridge, Ian Craddock, Ryan McConville, Raul Santos-Rodriguez, Kevin Chetty, and Robert Piechocki. Operanet, a multimodal activity recognition dataset acquired from radio frequency and vision-based sensors. *Scientific Data*, 9:474, 2022. doi: 10.1038/s41597-022-01573-2.

Xu Cao, Tong Zhou, Yunsheng Ma, Wenqian Ye, Can Cui, Kun Tang, Zhipeng Cao, Kaizhao Liang, Ziran Wang, James M. Rehg, and Chao Zheng. Maplm: A real-world large-scale vision-language benchmark for map and traffic scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21819–21830, 2024.

Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 168–172, 2015. doi: 10.1109/ICIP.2015.7350781.

Corentin Dancette, Remi Cadene, Damien Teney, and Matthieu Cord. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1574–1583, 2021.

Chenyou Fan. Egovqa: An egocentric video question answering benchmark dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 4359–4366, 2019. doi: 10.1109/ICCVW.2019.00536.

Gueter Josmy Faure, Min-Hung Chen, Jia-Fong Yeh, Hung-Ting Su, and Winston H. Hsu. FineBench: Benchmarking and enhancing vision-language models for fine-grained human activity understanding. *arXiv preprint arXiv:2605.19846*, 2026. doi: 10.48550/arXiv.2605.19846.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6904–6913, 2017. doi: 10.1109/CVPR.2017.670.

Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3343–3360, 2022. doi: 10.52202/068431-0242.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024.

Siyang Jiang, Mu Yuan, Xiang Ji, Bufang Yang, Zeyu Liu, Lilin Xu, Yang Li, Yuting He, Liran Dong, Wenrui Lu, Zhenyu Yan, Xiaofan Jiang, Wei Gao, Hongkai Chen, and Guoliang Xing. A large-scale multimodal dataset and benchmarks for human activity scene understanding and reasoning. In *Proceedings of the 24th Annual International Conference on Mobile Systems, Applications and Services*, MobiSys '26, pp. 352–370. Association for Computing Machinery, 2026. doi: 10.1145/3745756.3809209.

Quan Kong, Ziming Wu, Ziwei Deng, Martin Klinkigt, Bin Tong, and Tomokazu Murakami. Mmact: A large-scale dataset for cross modal human action understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8658–8667, 2019.

Yuxuan Li, Vijay Veerabadran, Michael L. Iuzzolino, Brett D. Roads, Asli Celikyilmaz, and Karl Ridgeway. EgoToM: Benchmarking theory of mind reasoning from egocentric videos. *arXiv preprint arXiv:2503.22152*, 2025.

Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. NTU RGB+D 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2684–2701, 2020. doi: 10.1109/TPAMI.2019.2916873.

Shaoyu Liu, Jianing Li, Guanghui Zhao, Yunjian Zhang, Xin Meng, Fei Richard Yu, Xiangyang Ji, and Ming Li. Eventgpt: Event stream understanding with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29139–29149, 2025.

Jiageng Mao, Junjie Ye, Yuxi Qian, Marco Pavone, and Yue Wang. A language agent for autonomous driving. In *Conference on Language Modeling (COLM)*, 2024.

Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, Elahe Arani, and Oleg Sinavski. Lingoqa: Visual question answering for autonomous driving. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 252–269, 2024. doi: 10.1007/978-3-031-72980-5_15.

Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2640–2649, October 2017. doi: 10.1109/ICCV.2017.288.

Mistral AI. Mistral-small-3.2-24b-instruct-2506. Hugging Face model release, 2025. Model: mistralai/Mistral-Small-3.2-24B-Instruct-2506.

Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. Berkeley mhad: A comprehensive multimodal human action database. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pp. 53–60, 2013. doi: 10.1109/WACV.2013.6474999.

Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4542–4550, 2024. doi: 10.1609/aaai.v38i5.28253.

Daniel Reich, Felix Putze, and Tanja Schultz. Measuring faithful and plausible visual grounding in vqa. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 3129–3144, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.206.

Ivan Rodin, Tz-Ying Wu, Kyle Min, Sharath Nittur Sridhar, Antonino Furnari, Subarna Tripathi, and Giovanni Maria Farinella. EASG-Bench: Video q&a benchmark with egocentric action scene graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 2732–2737, 2025.

Jeongwan Shin, Jaehyeon Kim, Donguk Ko, and Jaeho Choi. Can language models understand mmwave data? benchmarking large language models for mmwave radar-based human understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 2208–2219, 2026.

Ayush Shrivastava, Kirtan Gangani, Laksh Jain, Mayank Goel, and Nipun Batra. Thermeval: A structured benchmark for evaluation of vision-language models on thermal imagery. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26. Association for Computing Machinery, 2026. doi: 10.1145/3770855.3817469.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beisswenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. *arXiv preprint arXiv:2312.14150*, 2025.

Kexin Tian, Jingrui Mao, Yunlong Zhang, Jiwan Jiang, Yang Zhou, and Zhengzhong Tu. Nuscenes-spatialqa: A spatial understanding and reasoning benchmark for vision-language models in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 4626–4635, 2025.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

Xiao Wang, Jingyun Hua, Weihong Lin, Yuanxing Zhang, Fuzheng Zhang, Jianlong Wu, Di Zhang, and Liqiang Nie. HAIC: Improving human action understanding and generation with better captions for multi-modal large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10158–10181, Vienna, Austria, 2025b. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.501.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. Mmsi-bench: A benchmark for multi-image spatial intelligence. In *International Conference on Learning Representations (ICLR)*, 2026.

Dongsheng Yuan, Xie Zhang, Weiying Hou, Sheng Lyu, Yuemin Yu, Luca Jiang-Tao Yu, Chengxiao Li, and Chenshu Wu. Octonet: A large-scale multi-modal dataset for human activity understanding grounded in motion-captured 3d pose labels. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0469.

Yuhui Zhang, Yuchang Su, Yiming Liu, Xiaohan Wang, James Burgess, Elaine Sui, Chenyu Wang, Josiah Aklilu, Alejandro Lozano, Anjiang Wei, Ludwig Schmidt, and Serena Yeung-Levy. Automated generation of challenging multiple-choice questions for vision language model evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29580–29590, 2025. doi: 10.1109/CVPR52734.2025.02754.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4995–5004, 2016. doi: 10.1109/CVPR.2016.540.

A Appendix

A.1 Atomic Performance across Modalities

(a) All atomic questions.

Model	PC	E	T	R	All
Qwen2.5	32.5	28.4	58.4	38.0	39.9
Qwen3	32.5	49.8	64.5	48.0	48.9
Gemma-4	42.9	56.2	62.6	51.0	53.7
InternVL	23.6	44.8	51.4	44.0	40.2
Mistral	27.4	37.3	50.0	40.0	38.4

(b) Binary questions.

Model	PC	E	T	R
Qwen2.5	85.7	37.2	59.3	47.3
Qwen3	71.4	69.4	63.2	55.0
Gemma-4	92.9	62.1	65.4	52.3
InternVL	32.1	61.0	55.5	44.5
Mistral	35.7	56.6	52.6	39.1

(c) Quantitative questions.

Model	PC	E	T	R
Qwen2.5	30.2	10.0	–	31.3
Qwen3	29.6	10.0	–	31.3
Gemma-4	38.0	38.0	–	43.8
InternVL	17.9	20.0	–	31.3
Mistral	22.3	26.0	–	18.8

Table 7: Atomic performance by target modality. (a) Accuracy across all 687 atomic questions. (b) Balanced accuracy on binary questions. (c) Accuracy within the predefined tolerance on quantitative questions. PC, E, T, and R denote point cloud, event camera, thermal, and radar. The corresponding question counts are shown in Table 2. Thermal has no quantitative questions.

The 687 atomic questions differ in both target modality and answer format. Table 7 reports the full modality-level results, together with format-matched comparisons for binary and quantitative questions. Binary performance is reported using balanced accuracy to account for the distribution of yes and no answers, while quantitative performance uses the predefined tolerance associated with each question.

Across all 687 atomic questions, thermal questions obtain the highest accuracy for four of the five models, while Gemma-4 achieves the highest overall atomic accuracy. The modality results are not directly comparable because the answer-format distributions differ across modalities. The format-matched results show this difference. For example, Gemma-4 obtains 92.9% balanced accuracy on binary point-cloud questions but 38.0% accuracy on quantitative point-cloud questions.

A.2 Question Distribution across Activity and Visibility Categories

Table 8 reports the distribution of recordings and questions across the activity and visibility categories introduced in Sec. 3.1.

Table 8: Recorded activities and questions by activity and visibility category.

	Recordings	Atomic	Compound
Health & rehabilitation	28	293	91
Home & work	17	201	66
Exercise & sport	10	121	39
Social interaction	6	72	24
Clear / Occluded / Dark	47 / 9 / 5	572 / 82 / 33	180 / 32 / 8
Total	61	687	220

A.3 Quantitative Evaluation Tolerances

The 245 quantitative atomic questions are evaluated using tolerances defined before model evaluation. Distance questions use $\pm \max(0.05\,\mathrm{m}, 15\%$ of the ground-truth distance$)$. The 5 cm minimum accounts for LiDAR measurement precision and differences between observed body surfaces and motion-capture joint locations. The Livox Avia used in our setup has a reported range precision of 2 cm at 20 m. Angle questions use a tolerance of $\pm 10°$, consistent with the measurement variation reported for clinical joint-angle measurements. Time questions use ± 31.25 ms, corresponding to the temporal window over which events are accumulated into each visualization provided to the model. Rate and speed questions use $\pm \max(0.125\,\mathrm{m/s}, 15\%$ of the ground-truth speed$)$, where 0.125 m/s corresponds to one Doppler step in the radar representation.

Table 9: Scoring rules for the 245 quantitative atomic questions.

Quantity	Questions	Tolerance
Distance	159	$\pm \max(0.05\,\text{m}, 15\%$ of GT distance)
Angle	23	$\pm 10^\circ$
Time	50	$\pm 31.25\,\text{ms}$
Rate / speed	13	$\pm \max(0.125\,\text{m/s}, 15\%$ of GT speed)

A.4 Human Expert Verification

Human verification is conducted after the AI-assisted candidate generation described in Sec. 3.3. The total number of domain experts who participated in the process is five, and each candidate atomic question is reviewed by at least three researchers with prior publications related to the corresponding sensing modalities. Reviewers are provided with the recorded activity context and synchronized sensor observations.

For each atomic question, reviewers verify three properties: (1) the question is supported by the recorded activity, (2) the information required to answer the question is exclusive to the designated modality among the benchmark modalities, and (3) the information required to answer the question is provided by the designated modality rather than the other modalities. Questions that are ambiguous, unsupported by the recording, or answerable from an unintended modality are revised or removed.

Compound questions are reviewed after composition. Reviewers verify that each question is grounded in both constituent atoms, is applicable to the recorded activity, and preserves the information requirements of both designated modalities. Questions that do not preserve these properties after composition are revised or removed.

A.5 Sensor Specifications and Synchronization

Sensor setup. Data were recorded using four sensor rigs positioned around the capture area together with a Vicon optical motion-capture system providing the reference 3D human pose. Table 10 summarizes the sensors used for data collection.

Table 10: Sensor specifications used for SURE-Bench data collection.

Modality	Device	Resolution / output	Rate
Point cloud	Livox Avia LiDAR	$\sim 70^\circ \times 70^\circ$ FoV, $\sim$24k points/sweep	10 Hz
Event	iniVation DVXplorer	640×480, asynchronous events	–
Thermal	FLIR Boson	640×512	$\sim$20 Hz
Radar (Rig 1)	Altos Radar V2	3D point cloud with Doppler velocity	10 Hz
Radar (Rigs 2,3,4)	TI mmWave radar	Point measurements with Doppler velocity	10 Hz
RGB-D (reference only)	Intel RealSense D455	RGB 1280×720, depth 848×480	30 Hz
Motion capture	Vicon	Marker and segment poses	$\sim$200 Hz

Temporal synchronization. The sensor rigs record using independent clocks. We use the 10 Hz LiDAR stream from one reference rig as the temporal reference. For each LiDAR sweep, the temporally closest observation from the thermal, radar, RGB-D, and motion-capture streams is selected. Because the event-camera sensor timestamps have a fixed offset from the recording host, event observations are aligned using their host receive timestamps.

Model inputs. The observations provided to the model are selected according to the information required by each question. Thermal input consists of one frame. Event-camera input consists of 16 temporally ordered visualizations, each accumulating events over a 31.25 ms window, for a total duration of 0.5 s. Point-cloud input uses three LiDAR snapshots, and radar input uses five consecutive radar frames.