Using LLMs for Late Multimodal Sensor Fusion for Activity Recognition

Anonymous Author(s)
Affiliation
Address
email

Abstract

Sensor data streams provide valuable information around activities and context for downstream applications, though integrating complementary information can be challenging. We show that large language models (LLMs) can be used for late fusion for activity classification from audio and motion time series data. We curate a subset of data for diverse activity recognition across contexts (e.g., household activities, sports) from the Ego4D dataset. Evaluated LLMs achieve 12-class one-shot classification F1-scores significantly above chance, with no task-specific training. Zero-shot classification via LLM-based fusion from modality-specific models can enable multimodal temporal applications where there is limited aligned training data for directly learning a shared embedding space. Additionally, LLM-based fusion can enable model deploying without requiring additional memory and computation for targeted application-specific multimodal models.

1 Introduction and Related Work

We explore using large language models (LLM) for late fusion of time series modalities, to enable multimodal understanding directly from lightweight modality-specific models, without requiring additional training to learn a joint embedding space. To our knowledge, we are the first to explore reasoning capabilities of LLMs for late fusion of distinct multi-modal time series data streams for activity recognition. We focus on time series data because its ease of recording, storage, and privacy-sensitive characteristics make time series data streams a strong practical candidate for activity recognition – key context for many health applications. Late fusion via LLMs enables zero-shot learning with time series data, which is particularly impactful in the health domain where there is often limited training data and privacy is critical. Recognizing activity context can contribute to improved machine learning algorithms for health, where models may be impacted by unobserved contextual factors (e.g., interpreting heart rate data may be impacted by if a person is doing chores, socializing, or relaxing).

Extensive prior work has explored learning joint embedding spaces across modalities via a constrastive loss, and by combining constrastive and generative losses – including CLIP (Radford et al. [2021]), CLAP (Elizalde et al. [2023]), and recent works in the sensor and activity recognition domains including LLaSA, Sensor2Text, and SensorLM (Zheng and Arakawa [2025], Jiang et al. [2024], Imran et al. [2024], Chen et al. [2024], Hong et al. [2025]). For instance, Sensor2Text trains a Q-former fusion module while fine-tuning individual encoders to learn text representations of motion data and SensorLM employs a combination of contrastive and captioning losses to learn to describe trends in time series data. Capturing rich temporal and complementary semantics across modalities can be challenging. Learning a descriptive joint embedding space of time series signals requires large sample sizes and scaling to multiple modalities can be difficult because of the scarcity of aligned, paired data. For instance, models to fuse audio, motion, and/or text data often use trained

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Figure 1: Model Architecture for Prompt Creation.

modality-specific encoders to align audio-text pairs and sensor-text pairs (Zhang et al. [2025], Jiang et al. [2024], Imran et al. [2024]) and are limited by the availability of rich, descriptive captions for sensor data (Haresamudram et al. [2025]). Furthermore, such approaches do not readily poise a model to *integrate* information from multiple modalities when they are *complementary* — in fact, contrastive approaches may actively discourage learning relevant but orthogonal information.

Natively mutlimodal models like Gemini and GPT4o can process and respond to inputs from images, video, text, and audio but typically do not directly work with sensor data (e.g., motion data). Fine-tuning models to work with sensor data directly can enable their use for specific downstream tasks Kim et al. [2024], Chow et al. [2024]. For instance, HealthLLM uses prompting and fine-tuning to adapt LLMs for health prediction tasks like sleep and stress monitoring Kim et al. [2024]. Such approaches require tailored datasets and the resulting models cannot generalize to new sensor-based tasks.

By training on large, diverse text corpi, large language models (LLMs) learn general world knowledge that captures how modalities relate to each other and other contexts. Prior work has shown that human-in-the-loop reasoning around model predictions can improve performance Huang et al. [2024]. Prior work has also explored late fusion via LLMs for autonomous driving and other perception tasks (Ruan et al. [2025], Hou et al. [2025]) and for motion-related data streams in activity recognition (Post et al. [2025]). Recent advances in LLMs could similarly enable agent-in-the-loop systems for diverse modalities from multiple sensors for activity classification. Our contributions are: (1) the curation of a high-quality dataset for activity recognition from Ego4D, (2) benchmarking zero- and one-shot closed-set activity classification with LLMs and an initial investigation of open-ended activity classification, and (3) insights around current capabilities and future directions on strengths and weaknesses of modalities and LLM-based fusion via provided examples and ablation studies.

2 Methods

2.1 Dataset

We curated a dataset of day-to-day activities from the Ego4D dataset Grauman et al. [2022] by searching for activities of daily living within the provided narrative descriptions. The dataset includes 20 second samples from twelve high-level activities: vacuum cleaning, cooking, doing laundry, eating, playing basketball, playing soccer, playing with pets, reading a book, using a computer, washing dishes, watching TV, workout/weightlifting. These activities were selected to span a range of household and fitness tasks, and based on their prevalence in the larger dataset. The timestamp and video IDs of the segments used in our analysis are provided in the supplemental material, to enable replication of this task by others. While the evaluated dataset is small, it is high-quality – a high proportion of the larger dataset did not have IMU data, and information-rich segments for contextual activities are sparsely distributed throughout the dataset. In the analyzed subset, each segment was manually validated to ensure it contained the targeted activity. To our knowledge, it is one of the only readily curated sets of multi-modal audio and IMU data for assessing contextual activity recognition performance available.

2.2 Modeling

The exploration includes the following modalities:

- **Audio data**, processed using MS CLAP (Elizalde et al. [2023]) to generate audio captions and VGGish to generate audio labels Hershey et al. [2017]
- **Motion data** 3-axis accelerometer and 3-axis gyroscope from an inertial measurement unit (IMU), processed using an activity classification model trained on the target dataset
- **Extra Context (Synthetic)** Additional descriptive information like setting (e.g., indoor/outdoor) and heart rate zone, that could be captured from sensors. For our experiments, we synthetically generated this information in order to evaluate its utility in modality integration

Audio data is processed in 4 second windows and IMU data is processed in 2 second windows, using an overlap of 2 seconds. Our experiments utilize two Large Language Models (LLMs): a large, state-of-the-art model, Gemini-2.5-pro, and a smaller, high-performance model, DeepSeek-R1-Distill-Qwen-32B. Language models are prompted to classify the contextual activity given per-modality predictions for each time step (Figure 1). The full prompt and one-shot example are in Appendix C. We perform ablation studies to investigate the contribution of each modality, different user contexts, and the integration capabilities of the LLMs.

2.2.1 Closed-set evaluation

We evaluate performance zero-shot and few-shot classification in a closed-set environment. In this task, the list of high-level activities is appended at the end of the prompt (Appendix C) and instruct the model to select the most likely activity given the information in the prompt. We report accuracy and macro-F1 score. We conduct this experiment in two settings: a zero-shot setting, and a one-shot setting where an illustrative example (with illustrated reasoning) is pre-pended to the prompt to assess in-context learning capabilities.

2.2.2 Open-ended evaluation

A pre-defined list of possible activities may not always be available. To assess model performance in a more challenging open-ended scenario, we evaluated performance without including predefined list of candidate activities. Results include sample outputs for qualitative discussion. As a first quantitative evaluation, we introduce a second step where the `Qwen-32B` model maps the initial, unconstrained activity prediction to the best-fitting label from our original closed set. This evaluation has limitations since it introduces a second inference step and does not directly evaluate the quality of the open-ended output, but it provides an easily measurable metric around how close the open-ended response was to the actual class and enables understanding comparative performances.

3 Results and Discussion

The results for the one-shot closed-set evaluation are presented in Table 1. Zero-shot results are provided in Appendix Tables 3 and 4. The results of this open-ended inference task are summarized in Table 2. 95% confidence intervals were calculated via bootstrapping. Appendix B shows per-activity results for fusion via each LLM and Appendix D shows a selection of example temporal inputs and reasoning and prediction outputs.

Both models achieved classification results well above chance, or 8.3% accuracy given audio captions, labels, and activity predictions. Among the modality-specific models included, the audio caption predictions from CLAP were most informative towards overall activity predictions. Prediction accuracy using audio labels was impacted by inaccurate class predictions that heavily influenced reasoning (as in the second example in Appendix D, where the prediction of animal-related sounds incorrectly influenced the model prediction). Prediction accuracy using activity labels was similarly influenced by mispredictions by the modality specific-model – for instance, when minimal motion was recognized from recorded data, the model tended to predict stationary activities like 'eating' or 'reading a book'. The inclusion of synthetic context, which was designed to mimic information that could be interpreted or collected from sensors, improved performance compared to the predictions that used audio and IMU data only.

While the activity predictions did not necessarily improve aggregate performance, they were used by the model in reasoning correct predictions in some cases (as in the first example in Appendix D) and for some classes (e.g., high motion classes like 'playing soccer' as shown in Tables 5 and

Table 1: Closed-set high-level activity classification performance in the one-shot setting (95% CIs).

Model (Setting)	Modalities				Accuracy (%)	Macro-F1 (%)
	Audio Caption	Audio Labels	IMU Activity Pred.	Extra Context		
Gemini-2.5-pro (closed-set)	✓	✓	✓	✓	68 (59, 76)	66 (56, 74)
	✓	✓	✓	×	63 (55, 72)	62 (52, 70)
	✓	×	×	×	60 (50, 69)	60 (50, 67)
	×	✓	×	×	41 (32, 50)	35 (26, 43)
	×	×	✓	×	13 (8, 20)	10 (5, 14)
Qwen-32B (closed-set)	✓	✓	✓	✓	56 (47, 65)	56 (45, 64)
	✓	✓	✓	×	51 (43, 60)	51 (41, 57)
	✓	×	×	×	52 (43, 61)	52 (42, 60)
	×	✓	×	×	42 (33, 51)	36 (29, 42)
	×	×	✓	×	13 (8, 20)	8 (3, 12)

Table 2: Open-ended high-level activity classification performance in the one-shot setting (95% CIs).

Model (Setting)	Modalities				Accuracy (%)	Macro-F1 (%)
	Audio Caption	Audio Labels	IMU Activity Pred.	Extra Context		
Gemini-2.5-pro (open-ended)	✓	✓	✓	✓	58 (50, 68)	58 (49, 65)
	✓	✓	✓	×	54 (45, 63)	52 (42, 60)
	✓	×	×	×	53 (44, 63)	51 (42, 59)
	×	✓	×	×	39 (31, 48)	35 (26, 42)
	×	×	✓	×	8 (4, 13)	5 (2, 9)
Qwen-32B (open-ended)	✓	✓	✓	✓	51 (43, 59)	53 (43, 60)
	✓	✓	✓	×	46 (37, 55)	46 (36, 53)
	✓	×	×	×	53 (43, 61)	54 (45, 61)
	×	✓	×	×	34 (26, 43)	33 (25, 40)
	×	×	✓	×	13 (7, 19)	8 (3, 13)

4 in Appendix B). The activity classification model used in these experiments had a limited output space compared to the audio models. A more powerful and flexible motion-based model would likely further help disambiguate between high- and low- motion activities.

Performance was also significantly above chance in the open-ended evaluation experiments where no activity list was provided (Table 2), highlighting the promise of using LLM-based reasoning for modality fusion without requiring additional training for modality alignment. The third and fourth examples in Appendix D show the model's reasoning for inferring the context of cooking in open-ended cases, and the fifth example shows an example where the model incorrectly predicts playing tennis instead of basketball, based on the motion predictions and recognized ball sounds.

4 Conclusions and Future Work

We present a first analysis of LLM-based fusion for activity recognition from time series data. We show that LLMs can be used to fuse predictions from modality-specific models for activity recognition, without requiring additional training or modality alignment. Fusing time series modalities without requiring learning a joint embedding space is also particularly salient for data-scarce time series tasks in sensing and health. Sensor-based activity recognition also has the potential to enhance the deployment of health models by providing valuable context.

We include a curated list of segments from a public multimodal dataset, Ego4D, to enable replicating and building upon this work. The presented experiments included a limited number of samples, activities, and a limited output space for the IMU modality. Future work will expand on the evaluated data and modality-specific models, and also explore strategies for training LLMs for targeted reasoning skills (Muennighoff et al. [2025], Yue et al. [2024], Højer et al. [2025]) for modality integration.

References

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

Wenqi Zheng and Yutaka Arakawa. A review of wearables-based activities of daily living recognition with llms: Overview, progress and trends. *IEEE Access*, 2025.

Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. Neurolm: A universal multi-task foundation model for bridging the gap between language and eeg signals. *arXiv preprint arXiv:2409.00101*, 2024.

Sheikh Asif Imran, Mohammad Nur Hossain Khan, Subrata Biswas, and Bashima Islam. Llasa: A multimodal llm for human activity analysis through wearable and smartphone sensors. *arXiv preprint arXiv:2406.14498*, 2024.

Wenqiang Chen, Jiaxuan Cheng, Leyao Wang, Wei Zhao, and Wojciech Matusik. Sensor2text: Enabling natural language interactions for daily activity tracking using wearable sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(4):1–26, 2024.

Zhiqing Hong, Yiwei Song, Zelong Li, Anlan Yu, Shuxin Zhong, Yi Ding, Tian He, and Desheng Zhang. Llm4har: Generalizable on-device human activity recognition with pretrained llms. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 4511–4521, 2025.

Yuwei Zhang, Kumar Ayush, Siyuan Qiao, A Ali Heydari, Girish Narayanswamy, Maxwell A Xu, Ahmed A Metwally, Shawn Xu, Jake Garrison, Xuhai Xu, et al. Sensorlm: Learning the language of wearable sensors. *arXiv preprint arXiv:2506.09108*, 2025.

Harish Haresamudram, Apoorva Beedu, Mashfiqui Rabbi, Sankalita Saha, Irfan Essa, and Thomas Ploetz. Limitations in employing natural language supervision for sensor-based human activity recognition-and ways to overcome them. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 273–281, 2025.

Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. Health-llm: large language models for health prediction via wearable sensor data, arxiv. *arXiv preprint arXiv:2401.06866*, 2024.

Winnie Chow, Lauren Gardiner, Haraldur T Hallgrímsson, Maxwell A Xu, and Shirley You Ren. Towards time series reasoning with llms. *arXiv preprint arXiv:2409.11376*, 2024.

Jeremy Zhengqi Huang, Reyna Wood, Hriday Chhabria, and Dhruv Jain. A human-ai collaborative approach for designing sound awareness systems. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2024.

Shulan Ruan, Rongwei Wang, Xuchen Shen, Huijie Liu, Baihui Xiao, Jun Shi, Kun Zhang, Zhenya Huang, Yu Liu, Enhong Chen, et al. A survey of multi-sensor fusion perception for embodied ai: Background, methods, challenges and prospects. *arXiv preprint arXiv:2506.19769*, 2025.

Xinmeng Hou, Wuqi Wang, Long Yang, Hao Lin, Jinglun Feng, Haigen Min, and Xiangmo Zhao. Driveagent: Multi-agent structured reasoning with llm and multimodal sensor fusion for autonomous driving. *arXiv preprint arXiv:2505.02123*, 2025.

Kevin Post, Reo Kuchida, Mayowa Olapade, Zhigang Yin, Petteri Nurmi, and Huber Flores. Contextllm: Meaningful context reasoning from multi-sensor and multi-device data using llms. In *Proceedings of the 26th International Workshop on Mobile Computing Systems and Applications*, pages 13–18, 2025.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

Murong Yue, Wenlin Yao, Haitao Mi, Dian Yu, Ziyu Yao, and Dong Yu. Dots: Learning to reason dynamically in llms via optimal reasoning trajectories search. *arXiv preprint arXiv:2410.03864*, 2024.

Bertram Højer, Oliver Jarvis, and Stefan Heinrich. Improving reasoning performance in large language models via representation engineering. *arXiv preprint arXiv:2504.19483*, 2025.

Appendix A Zero-shot results for closed-set classification

Table 3: Closed-set high-level activity classification performance in the zero-shot setting (95% CIs calculated with bootstrapping).

Model	Modalities				Accuracy (%)	Macro-F1 (%)
	Audio Caption	Audio Labels	IMU Activity Pred.	Extra Context		
Gemini-2.5-pro	✓	✓	✓	✓	63 (55, 72)	61 (52, 68)
	✓	✓	✓	×	60 (52, 69)	59 (49, 66)
	✓	×	×	×	60 (51, 68)	59 (49, 66)
	×	✓	×	×	49 (41, 58)	45 (36, 52)
	×	×	✓	×	11 (6, 17)	7 (3, 10)
Qwen-32B	✓	✓	✓	✓	52 (43, 61)	52 (41, 60)
	✓	✓	✓	×	48 (40, 58)	47 (37, 55)
	✓	×	×	×	48 (40, 58)	48 (39, 56)
	×	✓	×	×	41 (33, 50)	38 (29, 45)
	×	×	✓	×	10 (5, 15)	4 (1, 6)

Appendix B Per-activity performance for closed-set classification

Table 5: Closed-set one-shot per-activity performance with `Qwen-32B`.

High-level Activity	Modalities: Audio Caption	Audio Labels	IMU Activity Pred.	Extra Context	Precision	Recall	F1
Vacuum Cleaning	✓	✓	✓	✓	0.21	0.50	0.29
	✓	✓	✓	×	0.21	0.50	0.29
	✓	×	×	×	0.20	0.40	0.27
	×	✓	×	×	0.33	0.30	0.32
	×	×	✓	×	0.12	1.00	0.21
Cooking	✓	✓	✓	✓	0.58	0.70	0.64
	✓	✓	✓	×	0.83	0.50	0.62
	✓	×	×	×	0.70	0.70	0.70
	×	✓	×	×	0.47	0.70	0.56
	×	×	✓	×	0.00	0.00	0.00
Doing laundry	✓	✓	✓	✓	1.00	0.30	0.46
	✓	✓	✓	×	0.20	0.10	0.13
	✓	×	×	×	1.00	0.20	0.33
	×	✓	×	×	0.00	0.00	0.00
	×	×	✓	×	0.00	0.00	0.00
Eating	✓	✓	✓	✓	1.00	0.50	0.67
	✓	✓	✓	×	0.71	0.50	0.59
	✓	×	×	×	0.75	0.30	0.43
	×	✓	×	×	0.50	0.70	0.58
	×	×	✓	×	0.00	0.00	0.00
Playing basketball	✓	✓	✓	✓	0.80	0.80	0.80
	✓	✓	✓	×	1.00	0.80	0.89
	✓	×	×	×	0.86	0.60	0.71
	×	✓	×	×	0.80	0.40	0.53
	×	×	✓	×	0.00	0.00	0.00
Playing soccer	✓	✓	✓	✓	0.80	0.40	0.53
	✓	✓	✓	×	1.00	0.30	0.46
	✓	×	×	×	0.40	0.20	0.27
	×	✓	×	×	0.00	0.00	0.00
	×	×	✓	×	0.60	0.30	0.40
Playing with pets	✓	✓	✓	✓	0.62	1.00	0.77
	✓	✓	✓	×	0.59	1.00	0.74
	✓	×	×	×	0.77	1.00	0.87
	×	✓	×	×	0.33	0.60	0.43
	×	×	✓	×	0.00	0.00	0.00
Reading a book	✓	✓	✓	✓	0.60	0.30	0.40
	✓	✓	✓	×	1.00	0.30	0.46
	✓	×	×	×	0.80	0.40	0.53
	×	✓	×	×	0.25	0.20	0.22
	×	×	✓	×	0.00	0.00	0.00
Using computer	✓	✓	✓	✓	0.50	0.70	0.58
	✓	✓	✓	×	0.53	0.80	0.64
	✓	×	×	×	0.40	0.60	0.48
	×	✓	×	×	0.29	0.50	0.37
	×	×	✓	×	0.00	0.00	0.00
Washing dishes	✓	✓	✓	✓	0.89	0.80	0.84
	✓	✓	✓	×	0.67	0.80	0.73
	✓	×	×	×	0.88	0.70	0.78
	×	✓	×	×	0.82	0.90	0.86
	×	×	✓	×	0.00	0.00	0.00
Watching TV	✓	✓	✓	✓	0.67	0.20	0.31
	✓	✓	✓	×	0.67	0.20	0.31
	✓	×	×	×	0.43	0.60	0.50
	×	✓	×	×	0.39	0.70	0.50
	×	×	✓	×	1.00	0.10	0.18
Workout/ Weightlifting	✓	✓	✓	✓	0.36	0.50	0.42
	✓	✓	✓	×	0.18	0.30	0.22
	✓	×	×	×	0.29	0.50	0.37
	×	✓	×	×	0.00	0.00	0.00
	×	×	✓	×	0.12	0.20	0.15

Table 4: Closed-set one-shot per-activity performance with `Gemini-2.5-pro`.

High-level Activity	Modalities				Precision	Recall	F1
	Audio Caption	Audio Labels	IMU Activity Pred.	Extra Context			
Vacuum Cleaning	✓	✓	✓	✓	0.50	0.40	0.44
	✓	✓	✓	×	0.38	0.30	0.33
	✓	×	×	×	0.33	0.30	0.32
	×	✓	×	×	1.00	0.20	0.33
	×	×	✓	×	0.20	0.10	0.13
Cooking	✓	✓	✓	✓	0.69	0.90	0.78
	✓	✓	✓	×	0.67	1.00	0.80
	✓	×	×	×	0.64	0.70	0.67
	×	✓	×	×	0.47	0.90	0.62
	×	×	✓	×	0.00	0.00	0.00
Doing laundry	✓	✓	✓	✓	0.80	0.40	0.53
	✓	✓	✓	×	0.80	0.40	0.53
	✓	×	×	×	0.62	0.50	0.56
	×	✓	×	×	0.00	0.00	0.00
	×	×	✓	×	0.20	0.30	0.24
Eating	✓	✓	✓	✓	0.86	0.60	0.71
	✓	✓	✓	×	0.88	0.70	0.78
	✓	×	×	×	1.00	0.70	0.82
	×	✓	×	×	0.39	0.70	0.50
	×	×	✓	×	0.00	0.00	0.00
Playing basketball	✓	✓	✓	✓	0.75	0.90	0.82
	✓	✓	✓	×	0.70	0.70	0.70
	✓	×	×	×	0.82	0.90	0.86
	×	✓	×	×	1.00	0.40	0.57
	×	×	✓	×	0.00	0.00	0.00
Playing soccer	✓	✓	✓	✓	1.00	0.40	0.57
	✓	✓	✓	×	1.00	0.30	0.46
	✓	×	×	×	0.80	0.40	0.53
	×	✓	×	×	0.00	0.00	0.00
	×	×	✓	×	0.67	0.40	0.50
Playing with pets	✓	✓	✓	✓	0.53	1.00	0.69
	✓	✓	✓	×	0.56	1.00	0.71
	✓	×	×	×	0.71	1.00	0.83
	×	✓	×	×	0.38	0.80	0.52
	×	×	✓	×	0.00	0.00	0.00
Reading a book	✓	✓	✓	✓	1.00	0.30	0.46
	✓	✓	✓	×	1.00	0.30	0.46
	✓	×	×	×	0.75	0.30	0.43
	×	✓	×	×	0.67	0.20	0.31
	×	×	✓	×	0.50	0.10	0.17
Using computer	✓	✓	✓	✓	0.67	0.80	0.73
	✓	✓	✓	×	0.73	0.80	0.76
	✓	×	×	×	0.64	0.70	0.67
	×	✓	×	×	0.62	0.80	0.70
	×	×	✓	×	0.00	0.00	0.00
Washing dishes	✓	✓	✓	✓	0.88	0.70	0.78
	✓	✓	✓	×	0.88	0.70	0.78
	✓	×	×	×	0.86	0.60	0.71
	×	✓	×	×	1.00	0.10	0.18
	×	×	✓	×	0.00	0.00	0.00
Watching TV	✓	✓	✓	✓	0.88	0.70	0.78
	✓	✓	✓	×	0.73	0.80	0.76
	✓	×	×	×	0.50	0.80	0.62
	×	✓	×	×	0.19	0.70	0.30
	×	×	✓	×	0.00	0.00	0.00
Workout/ Weightlifting	✓	✓	✓	✓	0.48	1.00	0.65
	✓	✓	✓	×	0.30	0.60	0.40
	✓	×	×	×	0.18	0.30	0.22
	×	✓	×	×	1.00	0.10	0.18
	×	×	✓	×	0.08	0.70	0.14

Appendix C LLM Prompts

Prompt for activity classification via late fusion

RESPOND IN ENGLISH ONLY. KEEP YOUR ANSWER CONCISE.
Below are temporally ordered 'audio captions' and 'top-5 audio labels' generated based on the sounds captured from the environment. Information under every single 'Time step' is based on 2-second audio recordings that follow each other.Audio information is accompanied by 'physical activity' predictions based on data recorded from a 'head-mounted' inertia measurement unit (IMU). At each 'Time step', we use a machine learning model to predict one of the following six labels: 'walking', 'running', 'standing', 'bend over', 'stand up', 'sit down'. When a predictions is 'Not available', it simply means that the model used for making that prediction (e.g., IMU data for activity prediction) was not available for that instance. For instance, if audio captions are 'Not available' for an instance, it means an audio captioning model was not available. It does NOT mean there was no sound. Or, if IMU model prediction is 'Not available' for an instance, it means an IMU physical activity prediction model was not available. It does NOT mean there was no or little physical activity.

Temporally ordered audio captions, audio labels, and IMU physical activity tags: {Per modality time series results for each time step}

IMPORTANT NOTE: The machine learning models that makes the audio and physical activity predictions are not perfect. They can make mistakes. For instance, the audio model can mistake the sound of wiping the windows for the sound of a DJ spinning the vinyl. Or, the physical activity prediction model can mistake that someone is running when they are walking fast. Therefore, do not hyperfocus on specific audio and physical activity tags. Focus on reasoning about the underlying sounds and physical motions that would lead to model predictions. Use information from across different timesteps and modalities to form a more robust big picture. In light of the information provided above, choose the most likely 'high-level activity' the camera wearer might be doing from. Think step by step and briefly explain your reasoning behind consistently combining the information across time steps. Finally, after reasoning, respond with the name of the activity at the end.

{For closed set evaluation} List of activities to choose from (please choose ONLY ONE and reply with the exact same name as in the list.): Playing soccer
Playing basketball
Cooking
Cleaning
Eating
Washing dishes
Doing laundry
Reading a book
Using computer
Watching TV
Workout/Weightlifting
Playing with pets

One shot example

I will give one illustrative example with physical activity and audio predictions, and some reasoning based on it to infer the high-level activity.
— Beginning of temporal audio and physical activity predictions for the illustrative example—
{— Time step 1 —} Top-5 audio labels with probabilities: Breaking: 0.51, Door: 0.15, Slam: 0.10, Speech: 0.03, Thunk: 0.02
Audio caption: Someone is banging pots and pans together
IMU model prediction: stand up
{— Time step 2 —} Top-5 audio labels with probabilities: Scrape: 0.22, Zipper (clothing): 0.17, Rub: 0.16, Wood: 0.07, Door: 0.03,
Audio caption: Someone is opening and closing a drawer.
IMU model prediction: running
{— Time step 3 —} Top-5 audio labels with probabilities: Rub: 0.62, Wood: 0.20, Sawing: 0.09, Sanding: 0.03, Chopping (food): 0.01,
Audio caption: Someone is cleaning a surface with a broom.
IMU model prediction: standing
{— Time step 4 —} Top-5 audio labels with probabilities: Rub: 0.24, Chopping (food): 0.12, Wood: 0.07, Sanding: 0.04, Tearing: 0.04,
Audio caption: Someone is cleaning a surface with a hammer.
IMU model prediction: walking
{— Time step 5 —} Top-5 audio labels with probabilities: Bathtub (filling or washing): 0.41, Water: 0.14, Water tap, faucet: 0.05, Speech: 0.04, Drip: 0.04,
Audio caption: A person is washing dishes in a sink.
IMU model prediction: walking
{— Time step 6 —} Top-5 audio labels with probabilities: Rub: 0.25, Gasp: 0.12, Speech: 0.10, Sound effect: 0.04, Snort: 0.03,
Audio caption: Something is rubbing against a surface
IMU model prediction: walking
{— Time step 7 —} Top-5 audio labels with probabilities: Door: 0.20, Water: 0.11, Slam: 0.09, Inside, small room: 0.08, Speech: 0.07,
Audio caption: Someone is washing dishes in a sink and then moving them around.
IMU model prediction: stand up
{— Time step 8 —} Top-5 audio labels with probabilities: Liquid: 0.21, Water: 0.19, Drip: 0.17, Inside, small room: 0.08, Speech: 0.05,
Audio caption: Water is splashing and clacking together as someone is washing dishes.
IMU model prediction: stand up
{— Time step 9 —} Top-5 audio labels with probabilities: Bathtub (filling or washing): 0.13, Drip: 0.12, Water tap, faucet: 0.12, Water: 0.12, Liquid: 0.08,
Audio caption: Water is splashing and splashing
IMU model prediction: stand up
— End of temporal audio and physical activity predictions for the illustrative example—
— Beginning of reasoning example for the illustrative example–
To determine the most likely high-level activity, I analyzed the temporal sequence of audio captions, audio labels, and IMU physical activity predictions across the 9 time steps. Each step represents 2 seconds of data, covering about 18 seconds in total. I focused on consistent patterns and underlying activities, considering the imperfection of the models (e.g., audio labels like "cleaning with a hammer" may be errors, and IMU predictions like "running" during drawer use are likely mismatches). Key insights:
- **Audio Captions and Labels Dominance:** From Time Step 5 onward (Steps 5–9), the audio captions explicitly describe dishwashing activities ("washing dishes in a sink," "moving them around," "water splashing and clacking"). This is supported by high-probability audio labels like "Bathtub (filling or washing)" (Step 5), "Liquid" and "Water" (Steps 7–9), which are characteristic of sink-based tasks. Earlier steps (1–4) include kitchen-like sounds (e.g., "banging pots and pans," "opening and closing a drawer," "cleaning a surface"), but these are precursors or related to the dishwashing theme (e.g., preparing or cleaning up).
- **IMU Physical Activity Integration:** The IMU predictions show frequent "stand up" (Steps 1, 7, 8, 9) and "walking" (Steps 4, 5, 6), which align with the physical motions of dishwashing—such as standing up from bending to retrieve items, walking to fetch or store dishes, or shifting position at the sink. "Standing" (Step 3) also fits. Outliers like "running" (Step 2) are likely model errors given the context (drawer opening doesn't involve running), so they were discounted.
- **Temporal Consistency:** The sequence evolves from general kitchen sounds (Steps 1–4) to sustained dishwashing focus (Steps 5–9), suggesting a single, cohesive activity like washing dishes. For instance: - Step 1 (stand up with pot banging) could be rising to start cleaning. - Steps 2–3 (drawer, cleaning with broom) might involve gathering tools or pre-cleaning. - Step 4's "cleaning with a hammer" is probably a misclassification (e.g., chopping food during prep, but dishwashing dominates later). - Steps 5–9 center on water, splashing, and dish movement, with repeated "stand up" indicating motions like lifting dishes or changing posture.
- **Activity Selection:** "Washing dishes" is the clearest match because: - It directly explains the water-related sounds and captions from Steps 5–9. - The IMU predictions (standing, walking, stand up) are consistent with the intermittent movements in dishwashing (e.g., standing at the sink, walking to put dishes away). - Kitchen elements (pots, drawers) from earlier steps are common in dishwashing contexts, as it often follows cooking or involves tidying.
Thus, "Washing dishes" best explains the consistent audio evidence and physical motions.
user activity: Washing dishes
— End of reasoning example for the illustrative example
Now, please come up with an answer for the original test example I included at the beginning.

Appendix D Selected examples

Correct prediction; Gemini; closed set (’Workout/weightlifting’)

Input — Beginning of temporal audio and physical activity predictions for the example—
{— Time step 1 —}
Top-5 audio labels with probabilities: Gasp: 0.56, Snort: 0.15, Speech: 0.04, Breathing: 0.03, Sneeze: 0.02,
Audio caption: A person is walking around and breathing heavily
IMU model prediction: stand up
{— Time step 2 —}
Top-5 audio labels with probabilities: Chewing, mastication: 0.20, Inside, small room: 0.12, Sound effect: 0.07, Speech: 0.04, Biting: 0.04,
Audio caption: A person is digging a hole in a hard surface.
IMU model prediction: stand up
{— Time step 3 —}
Top-5 audio labels with probabilities: Crunch: 0.29, Chewing, mastication: 0.14, Biting: 0.10, Inside, small room: 0.09, Sound effect: 0.05,
Audio caption: A person is digging through a hard surface
IMU model prediction: bend over
{— Time step 4 —}
Top-5 audio labels with probabilities: Gasp: 0.50, Snort: 0.10, Speech: 0.08, Sigh: 0.07, Inside, small room: 0.03 Audio caption: A person is breathing heavily and breathing heavily
IMU model prediction: stand up
{— Time step 5 —}
Top-5 audio labels with probabilities: Gasp: 0.70, Snort: 0.21, Speech: 0.02, Sigh: 0.01, Breathing: 0.01,
Audio caption: A person is breathing heavily and breathing heavily.
IMU model prediction: stand up
{— Time step 6 —}
Top-5 audio labels with probabilities: Sigh: 0.79, Gasp: 0.13, Snort: 0.02, Breathing: 0.01, Speech: 0.01,
Audio caption: A person is breathing heavily and repeatedly.
IMU model prediction: walking
{— Time step 7 —}
Top-5 audio labels with probabilities: Sigh: 0.53, Snort: 0.13, Gasp: 0.09, Breathing: 0.08, Sound effect: 0.03,
Audio caption: A person is walking slowly and carefully across a room.
IMU model prediction: walking
{— Time step 8 —}
Top-5 audio labels with probabilities: Snort: 0.44, Gasp: 0.15, Speech: 0.07, Sigh: 0.06, Animal: 0.02,
Audio caption: A person breathing heavily
IMU model prediction: bend over
{— Time step 9 —}
Top-5 audio labels with probabilities: Gasp: 0.26, Snort: 0.19, Speech: 0.08, Sneeze: 0.07, Breathing: 0.05,
Audio caption: A man coughs and exhales
IMU model prediction: bend over
— End of temporal audio and physical activity predictions —

Output
Step-by-Step Reasoning
To determine the most likely high-level activity, I analyzed the sequence of audio and IMU data. The key is to find a consistent theme across both modalities, discounting potential model errors.
1. **Audio Analysis**: The most dominant and recurring audio theme is strenuous breathing. Across nearly all time steps, the top audio labels include ‘Gasp‘, ‘Snort‘, and ‘Sigh‘, and the captions consistently describe "breathing heavily." This strongly indicates significant physical exertion. The sounds in steps 2 and 3 (‘Chewing‘, ‘Crunch‘, "digging") are likely misinterpretations of sounds related to effort, such as grunting, teeth clenching, or the sound of equipment.
2. **IMU Physical Activity Analysis**: The IMU data shows a clear pattern of dynamic movement: ‘stand up‘ (steps 1, 2, 4, 5), ‘bend over‘ (steps 3, 8, 9), and ‘walking‘ (steps 6, 7). This sequence of bending down, rising, and moving around is not random but suggests a cyclical, purposeful physical task.
3. **Synthesis and Integration**: Combining the audio and IMU data paints a clear picture. The heavy breathing (audio) is the direct result of the physical movements (IMU). The cycle of bending over, standing up, and walking is characteristic of many exercises. For example, one might bend to lift a weight (deadlift, kettlebell swing), stand up during the lift, and then walk or pace between sets.
4. **Activity Selection**: * Activities like ‘Cooking‘ or ‘Cleaning‘ would not typically produce such consistent and intense heavy breathing. * ‘Playing soccer‘ or ‘basketball‘ would likely involve more ‘running‘. * ‘Workout/Weightlifting‘ perfectly aligns with both the strenuous breathing sounds and the specific physical motions of bending, standing, and pacing recorded by the IMU.
Therefore, the combination of intense respiratory sounds and a repeated pattern of bending and standing strongly suggests a workout session.

user activity: Workout/Weightlifting

Incorrect prediction; Gemini; closed set ('Workout/weightlifting')

Input — Beginning of temporal audio and physical activity predictions for the example—
{— Time step 1 —}
Top-5 audio labels with probabilities: Animal: 0.13, Speech: 0.13, Inside, small room: 0.07, Walk, footsteps: 0.05, Clip-clop: 0.03,
Audio caption: A horse walks around
IMU model prediction: stand up
{— Time step 2 —}
Top-5 audio labels with probabilities: Tearing: 0.30, Zipper (clothing): 0.19, Inside, small room: 0.13, Speech: 0.05, Scrape: 0.02,
Audio caption: A person rubs a surface and then rubs a surface again.
IMU model prediction: stand up
{— Time step 3 —}
Top-5 audio labels with probabilities: Thunk: 0.15, Speech: 0.13, Inside, small room: 0.11, Breaking: 0.09, Door: 0.07,
Audio caption: A person is moving objects around.
IMU model prediction: stand up
{— Time step 4 —}
Top-5 audio labels with probabilities: Inside, small room: 0.19, Scrape: 0.10, Speech: 0.09, Gasp: 0.08, Snort: 0.05,
Audio caption: A person breathes and breathes
IMU model prediction: stand up
{— Time step 5 —}
Top-5 audio labels with probabilities: Inside, small room: 0.16, Biting: 0.11, Scrape: 0.07, Speech: 0.05, Door: 0.04,
Audio caption: A person is shuffling a piece of wood.
IMU model prediction: bend over
{— Time step 6 —}
Top-5 audio labels with probabilities: Snort: 0.51, Gasp: 0.26, Breathing: 0.03, Speech: 0.02, Sigh: 0.01,
Audio caption: A person rubs and breathes
IMU model prediction: bend over
{— Time step 7 —}
Top-5 audio labels with probabilities: Animal: 0.10, Sheep: 0.07, Livestock, farm animals, working animals: 0.06, Snort: 0.06, Speech: 0.05,
Audio caption: A person walks on a floor and a door opens.
IMU model prediction: bend over
{— Time step 8 —}
Top-5 audio labels with probabilities: Snort: 0.35, Gasp: 0.09, Speech: 0.06, Animal: 0.05, Inside, small room: 0.03,
Audio caption: A person walks slowly and then stops.
IMU model prediction: bend over
{— Time step 9 —}
Top-5 audio labels with probabilities: Snort: 0.34, Gasp: 0.14, Speech: 0.11, Animal: 0.04, Sigh: 0.02,
Audio caption: A person is walking and breathing heavily and then opens a door.
IMU model prediction: bend over
— End of temporal audio and physical activity predictions —

Output
Step-by-Step Reasoning To determine the most likely high-level activity, I analyzed the temporal sequence of audio and IMU data. I focused on identifying consistent patterns across both modalities while accounting for potential model inaccuracies.
1. **Dominant Audio Cues:** Across multiple time steps, the audio data strongly indicates the presence of an animal. Key labels include "Animal" (Steps 1, 7, 8, 9), "Clip-clop" (Step 1), "Livestock" (Step 7), and most notably, "Snort" which appears with high probability in the second half of the sequence (Steps 6, 7, 8, 9). While the initial caption "A horse walks around" is likely a misinterpretation for an indoor setting, the persistence of animal-related sounds is a crucial clue. Sounds like "rubbing," "shuffling," and heavy "breathing" or "gasping" accompany these animal sounds, suggesting a close interaction.
2. **Consistent Physical Motion:** The IMU data shows a clear pattern of movement. The person first performs a "stand up" motion (Steps 1-4) and then transitions into a sustained "bend over" posture for the remainder of the sequence (Steps 5-9). This physical sequence of standing and then bending over for a prolonged period aligns well with interacting with something at a low level, such as on the floor.
3. **Synthesizing Audio and IMU Data:** The combination of sustained bending over and the simultaneous sounds of an animal (snorting, gasping) and interaction (rubbing, shuffling) creates a coherent picture. The person likely stands up, approaches an animal, and then bends over to engage with it. The rubbing sounds could be petting, and the shuffling and breathing/snorting could be from active play.
4. **Activity Selection:** Comparing this synthesized scenario to the provided list, "Playing with pets" is the most logical fit. It explains the animal sounds, the interaction noises (rubbing), and the physical posture (bending over to be at the pet's level). Other activities like 'Cleaning' or 'Workout' might involve bending, but they do not account for the prominent and consistent animal-related audio cues.
user activity: Playing with pets

Correct prediction; Gemini; open ended (’Cooking’)

Input — Beginning of temporal audio and physical activity predictions for the example—
{— Time step 1 —}
Top-5 audio labels with probabilities: Water: 0.06, Sliding door: 0.06, Door: 0.06, Water tap, faucet: 0.04, Fill (with liquid): 0.04,
Audio caption: Food is frying and sizzling
IMU model prediction: walking
{— Time step 2 —}
Top-5 audio labels with probabilities: Speech: 0.07, Door: 0.05, Water: 0.03, Liquid: 0.03, Slam: 0.03,
Audio caption: A person is cooking food in a pan and then flipping it over.
IMU model prediction: walking
{— Time step 3 —}
Top-5 audio labels with probabilities: Speech: 0.39, Water: 0.04, Throat clearing: 0.04, Animal: 0.03, Music: 0.03,
Audio caption: A steady stream of water is splashing in a frying pan.
IMU model prediction: stand up
{— Time step 4 —}
Top-5 audio labels with probabilities: Squeal: 0.10, Vehicle: 0.07, Water: 0.07, Speech: 0.03, Mechanisms: 0.03,
Audio caption: Water is flowing in a faucet and splashing in the background.
IMU model prediction: stand up
{— Time step 5 —}
Top-5 audio labels with probabilities: Water: 0.11, Gush: 0.07, Speech: 0.06, Sizzle: 0.05, Water tap, faucet: 0.04,
Audio caption: A frying pan is frying in a pan.
IMU model prediction: stand up
{— Time step 6 —}
Top-5 audio labels with probabilities: Speech: 0.09, Gush: 0.07, Water tap, faucet: 0.06, Water: 0.06, Sink (filling or washing): 0.06,
Audio caption: A person is frying food in a pan.
IMU model prediction: bend over
{— Time step 7 —}
Top-5 audio labels with probabilities: Door: 0.13, Slam: 0.07, Dishes, pots, and pans: 0.07, Whack, thwack: 0.05, Speech: 0.05,
Audio caption: Something is fried and something is tapped
IMU model prediction: walking
{— Time step 8 —}
Top-5 audio labels with probabilities: Dishes, pots, and pans: 0.18, Door: 0.07, Inside, small room: 0.05, Cutlery, silverware: 0.05, Creak: 0.04,
Audio caption: Someone is pounding on a hard surface.
IMU model prediction: bend over
{— Time step 9 —}
Top-5 audio labels with probabilities: Slap, smack: 0.18, Finger snapping: 0.13, Speech: 0.06, Ping: 0.06, Door: 0.02,
Audio caption: Someone is tapping a hard object against a hard surface.
IMU model prediction: bend over
— End of temporal audio and physical activity predictions —

Output
Step-by-Step Reasoning
To determine the most likely high-level activity, I analyzed the sequence of audio and physical activity data. The analysis focuses on identifying consistent patterns that link the sounds and movements into a coherent event.
1. **Dominant Audio Theme:** The audio data consistently points to a kitchen environment and the act of cooking. Audio captions repeatedly mention "frying and sizzling" (Step 1), "cooking food in a pan" (Steps 2, 6), and "frying pan" (Steps 3, 5). This is strongly supported by audio labels like "Sizzle" (Step 5), "Dishes, pots, and pans" (Steps 7, 8), and "Water tap, faucet" (Steps 1, 5, 6), which are all characteristic of food preparation.
2. **Integration of Physical Activity:** The IMU predictions align well with the motions involved in cooking. * ‘walking‘ (Steps 1, 2, 7) corresponds to moving around a kitchen, for instance, between the stove, counter, and refrigerator. * ‘stand up‘ (Steps 3, 4, 5) is consistent with standing at a stove or counter to monitor or stir food. * ‘bend over‘ (Steps 6, 8, 9) fits with actions like getting items from a low cabinet, checking an oven, or leaning over a cutting board. The combination of "pounding" and "tapping" sounds with ‘bend over‘ (Steps 8, 9) strongly suggests preparing ingredients on a countertop.
3. **Temporal Consistency:** The sequence of events tells a logical story of preparing a meal. The activity starts with frying (Steps 1-2), involves adding water or using a faucet (Steps 3-4), and continues with frying (Steps 5-6). The final steps include sounds of utensils and dishes ("tapping," "pounding," "Dishes, pots, and pans"), which could be part of the final preparation or plating process.
4. **Activity Selection:** The combined evidence from both audio and motion sensors overwhelmingly supports "Cooking." The primary sounds are of frying and food preparation, not cleaning or eating. The physical movements are all typical for a person actively engaged in preparing a meal in a kitchen.
user activity: Cooking

Mostly correct prediction; Gemini; open ended (’Cooking’)

Input — Beginning of temporal audio and physical activity predictions for the example—
{— Time step 1 —}
Top-5 audio labels with probabilities: Vehicle: 0.19, Car: 0.08, Speech: 0.05, Water: 0.03, Medium engine (mid frequency): 0.03,
Audio caption: A steady stream of sizzling sizzles in the background.
IMU model prediction: walking
{— Time step 2 —}
Top-5 audio labels with probabilities: Sliding door: 0.21, Door: 0.17, Vehicle: 0.09, Slam: 0.07, Speech: 0.07,
Audio caption: A person is frying food in a pan and then sizzling it.
IMU model prediction: walking
{— Time step 3 —}
Top-5 audio labels with probabilities: Speech: 0.40, Door: 0.22, Sliding door: 0.06, Vehicle: 0.05, Cupboard open or close: 0.05,
Audio caption: A woman is running a heavy machine and then a woman is speaking.
IMU model prediction: stand up
{— Time step 4 —}
Top-5 audio labels with probabilities: Vehicle: 0.09, Car: 0.07, Speech: 0.05, White noise: 0.05, Water: 0.04,
Audio caption: A machine is running in the background.
IMU model prediction: stand up
{— Time step 5 —}
Top-5 audio labels with probabilities: Door: 0.20, Car: 0.16, Vehicle: 0.15, Sliding door: 0.13, Speech: 0.04,
Audio caption: A motor is running and a motor is moving
IMU model prediction: stand up
{— Time step 6 —}
Top-5 audio labels with probabilities: Vehicle: 0.22, Sliding door: 0.06, Speech: 0.05, Door: 0.04, Car: 0.04,
Audio caption: A car engine is running and rattling.
IMU model prediction: stand up
{— Time step 7 —}
Top-5 audio labels with probabilities: Speech: 0.17, Vehicle: 0.10, Water: 0.08, Slosh: 0.08, Music: 0.07,
Audio caption: A machine is running and a car is driving.
IMU model prediction: walking
{— Time step 8 —}
Top-5 audio labels with probabilities: Frying (food): 0.32, Sizzle: 0.25, Stir: 0.21, Vehicle: 0.05, Speech: 0.02,
Audio caption: A person is frying food in a pan and then sizzling it.
IMU model prediction: walking
{— Time step 9 —}
Top-5 audio labels with probabilities: Breaking: 0.16, Vehicle: 0.09, Scrape: 0.08, Glass: 0.06, Chink, clink: 0.06,
Audio caption: A metal object is being clattered against a metal surface.
IMU model prediction: walking
— End of temporal audio and physical activity predictions —

Output
Step-by-Step Reasoning
To determine the most likely high-level activity, I analyzed the consistent patterns across the audio and physical activity data.
1. **Dominant Audio Themes:** Two primary soundscapes emerge from the data. The first is clearly related to cooking, with highly specific labels and captions like "Frying (food)," "Sizzle," and "Stir" appearing multiple times (Steps 1, 2, 8). The second is a persistent background noise identified as a vehicle, with labels like "Vehicle," "Car," and captions mentioning a "motor" or "engine" in nearly every time step (1, 3, 4, 5, 6, 7).
2. **IMU Physical Activity Integration:** The physical activity consists of ‘walking‘ (Steps 1, 2, 7, 8, 9) and ‘stand up‘ (Steps 3, 4, 5, 6). This pattern of intermittent walking and shifting posture is consistent with moving around a limited workspace, such as a kitchen or a stall, rather than continuous walking or standing still.
3. **Temporal Consistency and Synthesis:** The cooking and vehicle sounds occur simultaneously throughout the sequence. For example, Step 2 mentions frying food while vehicle sounds are also detected. Step 8 again highlights frying and sizzling, while the surrounding steps are dominated by vehicle/engine sounds. This co-occurrence suggests the two activities are happening in the same location. The most logical explanation that combines active cooking with the constant sound of a running engine is that the cooking is taking place within or adjacent to a vehicle designed for this purpose. The sounds of a "Sliding door" (Steps 2, 3, 5, 6) could plausibly be the service window of such a vehicle.
4. **Activity Selection:** While "cooking outdoors" is a possibility, "Cooking in a food truck" provides a more precise explanation for the persistent and prominent "engine" and "motor" sounds described in the captions. This activity perfectly reconciles the specific cooking actions with the constant vehicle noise and the physical movements of working in a confined space.
user activity: Cooking in a food truck

Incorrect prediction; Gemini; open ended ('Playing basketball')

Input — Beginning of temporal audio and physical activity predictions for the example—
{— Time step 1 —}
Top-5 audio labels with probabilities: Speech: 0.12, Gasp: 0.10, Animal: 0.07, Chop: 0.07, Snort: 0.04,
Audio caption: Someone is walking on a hard surface.
IMU model prediction: stand up
{— Time step 2 —}
Top-5 audio labels with probabilities: Speech: 0.11, Vehicle: 0.07, Car: 0.06, Door: 0.05, Power windows, electric windows: 0.04,
Audio caption: A person is running and walking in a gravel pit.
IMU model prediction: running
{— Time step 3 —}
Top-5 audio labels with probabilities: Slam: 0.25, Speech: 0.06, Basketball bounce: 0.05, Whack, thwack: 0.04, Ping: 0.03,
Audio caption: A ball is bouncing and a person is grunting.
IMU model prediction: bend over
{— Time step 4 —}
Top-5 audio labels with probabilities: Speech: 0.14, Door: 0.10, Slam: 0.07, Gasp: 0.04, Chop: 0.04,
Audio caption: A rhythmic rhythmic beat is occurring in the background.
IMU model prediction: stand up
{— Time step 5 —}
Top-5 audio labels with probabilities: Chop: 0.65, Speech: 0.03, Whack, thwack: 0.02, Wood: 0.02, Tap: 0.01,
Audio caption: A person is walking while a stream of water is flowing.
IMU model prediction: stand up
{— Time step 6 —}
Top-5 audio labels with probabilities: Drum: 0.18, Percussion: 0.13, Musical instrument: 0.13, Music: 0.13, Slam: 0.04,
Audio caption: A man is walking and a loud banging sound
IMU model prediction: stand up
{— Time step 7 —}
Top-5 audio labels with probabilities: Door: 0.12, Music: 0.12, Sliding door: 0.08, Slam: 0.07, Speech: 0.06,
Audio caption: A ball is being bounced around in a large room.
IMU model prediction: stand up
{— Time step 8 —}
Top-5 audio labels with probabilities: Speech: 0.14, Animal: 0.09, Basketball bounce: 0.04, Tap: 0.03, Slap, smack: 0.03,
Audio caption: A ball bouncing in the background
IMU model prediction: stand up
{— Time step 9 —}
Top-5 audio labels with probabilities: Speech: 0.23, Animal: 0.16, Basketball bounce: 0.11, Clip-clop: 0.06, Horse: 0.04,
Audio caption: A tennis ball is bouncing back and forth in the background.
IMU model prediction: bend over
— End of temporal audio and physical activity predictions —

Output
Step-by-Step Reasoning
To determine the most likely high-level activity, I analyzed the temporal sequence of audio and IMU data. The analysis focuses on identifying consistent patterns across modalities and time steps.
1. **Dominant Audio Theme: Ball Activity:** A recurring theme across multiple time steps is the sound of a ball. This is explicitly mentioned in the audio captions for steps 3 ("A ball is bouncing"), 7 ("A ball is being bounced around"), 8 ("A ball bouncing"), and 9 ("A tennis ball is bouncing back and forth"). This is further supported by the high-probability audio label "Basketball bounce" appearing in steps 3, 8, and 9.
2. **Dominant Audio Theme: Impact Sounds:** Several time steps feature sharp, percussive sounds consistent with a sport. These include labels like "Slam" (steps 3, 4, 6), "Whack, thwack" (step 3), and "Chop" (step 5, with a high probability of 0.65). These sounds are likely the impact of a ball being hit, either with a racket or against a surface.
3. **IMU Physical Activity Integration:** The IMU predictions show a highly active user, inconsistent with a sedentary activity. The sequence includes 'running' (step 2), 'bend over' (steps 3 and 9), and frequent 'stand up' motions. This combination of running, bending, and recovering posture strongly suggests athletic movements, such as chasing a ball, getting into a low stance to hit it, and then standing up. The "grunt" mentioned in the caption for step 3 also indicates physical exertion.
4. **Synthesizing the Evidence:** The combination of continuous ball bouncing sounds, sharp impact noises ("whack," "chop," "slam"), and dynamic physical movements (running, bending) strongly points to a ball sport. The caption in step 9, "A tennis ball is bouncing back and forth," provides the most specific clue. The "back and forth" motion is characteristic of a racket sport like tennis, rather than dribbling in basketball. The "Whack" and "Chop" sounds are also more indicative of a racket striking a ball. The movements of running and bending are fundamental to playing tennis.
Therefore, the evidence consistently points to the user playing a racket sport, with tennis being the most likely candidate.
user activity: Playing tennis