EXTENDED-ABSTRACT

MultiSurf-GPT: Facilitating Context-Aware Reasoning with Large-Scale Language Model Agents for Multimodal Surface Sensing

YONGQUAN HU, UNSW Sydney, Sydney, NSW, Australia

BLACK SUN, Southern University of Science and Technology, Shenzhen, Guangdong, China

PENGCHENG AN, Southern University of Science and Technology, Shenzhen, Guangdong, China

ZHUYING LI, Southeast University, Nanjing, Jiangsu, China

WENCHEN HU, UNSW Sydney, Sydney, NSW, Australia

AARON JOHN QUIGLEY, Commonwealth Scientific and Industrial Research Organisation, Canberra, ACT, Australia

MultiSurf-GPT: Facilitating Context-Aware Reasoning with Large-Scale Language Models for Multimodal Surface Sensing

Yongquan Hu
yongquan.hu@unsw.edu.au
University of New South Wales
Sydney, NSW, Australia

Black Sun
blackthompson770@gmail.com
Southern University of Science and
Technology
Shenzhen, China

Pengcheng An
anpengcheng88@gmail.com
Southern University of Science and
Technology
Shenzhen, China

Zhuying Li
yongquan.hu@unsw.edu.au
University of New South Wales
Sydney, NSW, Australia

Wen Hu
wen.hu@unsw.edu.au
University of New South Wales
Sydney, NSW, Australia

Aaron J. Quigley
aquigley@acm.org
CSIRO's Data61
Sydney, NSW, Australia

ABSTRACT

Surface sensing is widely employed in health diagnostics, manufacturing and safety monitoring. Advances in mobile sensing affords this potential for context awareness in mobile computing, typically with a single sensing modality. Emerging multimodal large-scale language models offer new opportunities. We propose MultiSurf-GPT, which utilizes the advanced capabilities of GPT-4o to process and interpret diverse modalities (radar, microscope and multispectral data) uniformly based on prompting strategies (zero-shot and few-shot prompting). We preliminarily validated our framework by using MultiSurf-GPT to identify low-level information, and to infer high-level context-aware analytics, demonstrating the capability of augmenting context-aware insights. This framework shows promise as a tool to expedite the development of more complex context-aware applications in the future, providing a faster, more cost-effective, and integrated solution.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing**; • **Ubiquitous and mobile computing systems and tools**;

KEYWORDS

Surface Sensing, Context Aware, Large Language Model, Prompt Engineering.

1 INTRODUCTION

Ubiquitous and mobile computing fundamentally transforms our interaction with the digital world, where context awareness plays a pivotal role in enriching user experience and improving operational efficiency [1, 4, 26]. Particularly, surface sensing — the capability to discern surface materials or categorize objects — is essential for understanding the immediate environment of a mobile device [12, 24, 35]. This understanding aids devices in adjusting their functionalities based on context, whether optimizing screen readability based on ambient light conditions or human activity states [14, 32] or adjusting the interface when placed in different locations [7].

Despite advancements, a significant research gap exists in the unified processing of multimodal surface sensing data. Current methodologies often focus on specific types of data derived from different electromagnetic wave bands, such as visible light image [15], radar signal in the radio band [10, 29] and multispectral data including infrared and ultraviolet light [23, 30]. Each method typically develops customized algorithms tailored for one modality without leveraging the potential synergies between them. This siloed approach limits the ability to process complex datasets where multiple modalities overlap, presenting a clear opportunity for a more integrated solution [11, 22].

Furthermore, traditional Artificial Intelligence (AI) methods, such as machine learning algorithms, for context awareness are typically confined to specific low-level tasks such as object recognition or material identification [2, 17]. However, these methods often lack the capability to perform high-level reasoning tasks that necessitate integrating various sensory inputs to offer comprehensive insights into a user's environment or behavior [3, 16]. Multimodal Large Language Models (LLMs) present new opportunities to amalgamate multi-dimensional surface sensory data, enhancing overall context awareness [9]. LLMs excel in synthesizing information from diverse sources and generating rich context insights, predominantly based on processing and assimilating extensive generalized world knowledge [9]. It is crucial to note that despite the propensity of LLMs to generate insights with potential biases and inaccuracies, their capacity for logical and causal reasoning remains impressive [8, 33]. As technology evolves, these inherent risks are expected to diminish, which falls outside the purview of this study.

To address these challenges, we introduce MultiSurf-GPT, a framework leveraging the latest GPT-4o model to process multimodal surface sensing data (radar, microscope and multispectral data), aiming to establish a unified workflow from recognition to reasoning for high-level context awareness. Notably, the current capabilities of LLMs are primarily tailored to text, image, and file modalities [1]. Currently, how to uniformly process multimodal sensor data using LLMs remains an open question. Optimizing LLMs' recognition of sensor data typically requires instruction tuning, which is resource-intensive [21]. Alternatively, converting all data to the native modalities supported by GPT-40, such as images or CSV files, and using the LLM as an interface to invoke existing classification algorithms based on prompt engineering provides a cost-effective and efficient solution for rapid prototyping [5, 20]. In this study, we adopt the prompt-based strategies (zero-shot prompting and prompting) for rapid prototyping, but we anticipate incorporating instruction fine-tuning in future work.

The main contributions of this paper are threefold: (1) We introduce MultiSurf-GPT, using GPT-4o as an agent within a unified 2R framework for enhanced contextual information reasoning. We consider it as a tool to advance the future development of more adaptive and intelligent mobile applications; (2) We demonstrate the application of LLMs in the unified processing of multimodal surface sensing data, with our test results establishing a benchmark for further work, including case studies on high-level reasoning from diverse contexts; (3) We briefly dicuss the limitations and future directions, providing insights for ongoing efforts. Overall, this research not only broadens the technological landscape but also opens new avenues for the practical application of LLMs in ubiquitous computing.

2 RELATED WORK

2.1 Multimodal Surface Sensing Data for Ccontext Awareness

The analysis of multimodal surface sensing data is key to context awareness. With advancements in AI, surface sensing has become miniaturized and mobile. Based on electromagnetic band distribution, surface sensing data can be categorized into image (visible light band), multispectral (usually a combination of multiple invisible light bands such as infrared and ultraviolet), and radar (radio band). Firstly, visible images, such as the grayscale images used by Yang et al. in MagicFinger [28] or the RGB images used by Hu et al. in MicroCam [15], offer high resolution for detailed observation at a micro or nano scale, enhancing interpretability and expanding interactive possibilities. However, they suffer from a limited field of view and require extensive sample preparation. Secondly, multispectral data excels in differentiating materials based on spectral signatures, but it requires complex analysis and high equipment costs, making it uncommon on mobile devices. For example, Harrison et al. used wearable multispectral hardware for material classification [12], and Yeo et al. achieved high recognition accuracy with this sensor on a mobile device in SpeCam [30]. Lastly, radar data can penetrate obstacles and cover large areas, useful in wide applications such as meteorology and surveillance, but usually offers lower resolution

compared to optical methods and requires a clear line-of-sight. For instance, Yeo et al. used Google's Soli sensing suite in RadarCat for surface material classification [29]. Yeo et al. further extended radar-based context awareness to the recognition of tangible surfaces in Tangible Radar [31]. Additionally, Samitha et al. employed higher resolution radar signals in RadarFoot for fine-grained ground recognition and classification. [10]. In summary, current methods for context-aware surface sensing primarily concentrate on processing data from a single modality, with limited efforts directed towards the comprehensive understanding of multimodal data to deliver intricate contextual information about the environment.

2.2 LLM Applications in Sensing Data Processing and Reasoning

Recent advances in multimodal LLMs have significantly enhanced their capability to process sensor data, catalyzing a transformative shift across various domains, including environmental monitoring and autonomous systems. A prime example is EarthGPT, which integrates multimodal data streams such as optical, Synthetic Aperture Radar (SAR), and infrared data to improve the understanding of remote sensing images [34]. Similarly, VLMRemote utilizes multimodal LLMs to merge visual and linguistic modalities in remote sensing, enhancing geographic feature recognition by combining textual descriptions with visual data [19]. Moreover, multimodal LLMs demonstrate robust reasoning abilities over low-level sensor data. The LLMSense framework, for example, advances high-level reasoning over spatiotemporal sensor trajectories by leveraging multimodal LLMs' inherent capacity to assimilate extensive world knowledge and reasoning capabilities, effectively interpreting long-term environmental and situational dynamics [21]. HealthLLM employs multimodal LLMs to analyze health-related sensor data, diagnosing and predicting health conditions by integrating physiological data with contextual information for actionable healthcare insights [18]. In short, these developments highlight the potential of multimodal LLMs in the unified processing of multimodal sensor data. Their ability to interpret and reason about high-level information from low-level data is crucial for managing complex sensor inputs and paves the way for the development of smarter, more adaptive technologies.

3 METHODOLOGY

3.1 Dataset Selection

For context-aware surface sensing in mobile computing, we selected three datasets of different modalities: (1) Tangible Radar [31] dataset (radar data as CSV file): This dataset consists of radar signal recordings used to detect and classify physical objects within an environment. This dataset comprises radar reflection signals collected by a single-mode, multi-channel (8 channels, 2 transmitters, and 4 receivers) system. These signals encompass information about an object's distance, thickness, shape, density, and internal composition. (2) MicroCam [15] dataset (microscope image): This dataset comprises high-resolution images captured using microscopic cameras with the phone face-up, specifically designed to detail the textural and material properties of various surfaces. It is a valuable resource for researchers focusing on fine-grained image analysis and material identification. (3) SpeCam [30] dataset

(multispectral image): The data contains reflection light images, collected by placing a smartphone screen facing down and rapidly flashing different colors on the screen while the front camera captures the reflected light based on spectral cameras. This reflected light contains information about the surface material's color and optical properties.

3.2 Prompt Design

For our methods in MultiSurf-GPT framework, we utilize prompt engineering to address multiple tasks across various datasets as an initial exploration. These prompts are model-agnostic, and detailed descriptions of the language models and experimental settings will be provided in the following section. Our prompts are crafted to handle different modalities and ensure flexibility. Additionally, we have assessed and compared the effectiveness of zero-shot and prompts in our evaluations.

It is worth emphasizing that the processing of MultiSurf-GPT is not confined to specific datasets. After constructing a model to analyze a dataset, MultiSurf-GPT also review the description document (published papers) associated with it, extracting critical information from these documents, such as the sensing methods and the usage domains. Related details could be found in the "Extra settings of all datasets" part of the Section 4.

Zero-shot prompting. As shown in Table 1, the zero-shot prompting strategy consists of a specific task explaining, an additional rule to avoid unnecessary output, and restricted models to focus on the current task. Therefore, the final prompt for the model consisted of {Task Explaining} + {Rules}.

1-shot prompting. The prompt added the samples after the same zero-shot prompt template (We only use a one-shot prompt). Specifically, we include the task explaining prompt following the zero-shot prompt but provide the correct class labels instead of offering different candidate class labels for prediction.

4 EXPERIMENT

4.1 Settings

4.1.1 Dataset Settings. We leverage several input settings for different modalities data: (1) Tangible Radar dataset settings: we input a CSV file containing radar signals and class labels. This utilizes GPT-4o's ability to read CSV files (file modality) and employ the code interpreter to analyze the data using Python code for methods such as SVM [13]. (2) MircoCam dataset settings [15]: we input microscope images (image modality), utilizing GPT-4o's image modality recognition capabilities. For prompts, we provide one example image per category for GPT-4o to recognize. In contrast, no examples are given for zero-shot prompts. (3) SpeCam dataset settings [30]: we input multispectral images (image modality), again leveraging GPT-4o's image modality recognition abilities. Like the MircoCam settings, we provide one example image per category for prompts and none for zero-shot prompts. (4) Extra Settings of all datasets: the original research papers of the three datasets are also inputted as files, utilizing GPT-4o's capabilities to retrieve, read, and comprehend the documents.

4.1.2 Model Settings. Given that GPT-4o is the most recent series of GPT-4 that naturally supports multimodal capabilities, we adopted

GPT-4o as the tested LLMs. For the experiment, we tested the learning scenario to examine the capability of multi-model LLMs with limited information provided. We randomly selected one sample from the corresponding dataset as the sample in each repeated trial. We use the image updating module of GPT-4o. However, we use no other additional techniques (e.g., Chain-of-Thoughts [25]) to serve as a preliminary study of how multimodal LLMs process multimodal information. This approach ensures the results reflect the basic capability of the models, which was also consistent with previous work [27].

4.2 Results and Discussions

4.2.1 Performance of MultiSurf-GPT. Regarding the Tangible Radar dataset, MultiSurf-GPT demonstrated remarkable proficiency in generating machine learning code to analyze CSV data. Specifically, the tasks of Rotation, Orientation, and Identification achieved impressive accuracy rates of 100%, 90.91%, and 90.45%, respectively. However, Count, Order, and Distance tasks exhibited comparatively lower performance. This discrepancy is attributed to the inferior performance of the Random Forest (RF) model used by the code interpreter rather than the inherent capabilities of GPT-4o itself. For the MicroCam and SpeCam datasets, GPT-4o leveraged its image recognition capabilities. In zero-shot scenarios, GPT-4o struggled to identify images due to their limited semantic information accurately. Nonetheless, because of the high similarity between images, the accuracy of GPT-4o's recognition improved significantly in one-shot scenarios. Specifically, for the MicroCam dataset, the accuracy in Object and Material tasks increased by 23.33% and 11.11%, respectively. For the SpeCam dataset, the accuracy rose impressively by 51.67%. Moreover, MultiSurf-GPT excels in extracting detailed information from original research papers, such as the sensing methods and usage domains. These capabilities establish a robust foundation for more complex, context-aware analyses in future applications.

4.2.2 Case Analysis of Context-Awareness. As shown in Figure 2, the output is generated separately by GPT-4o and MultiSurf-GPT. As Dey et al. mentioned [1], context awareness encompasses four aspects: location, activity, time, and identity. The original GPT-4o provided overly simplistic answers and failed to address these four aspects adequately. In contrast, MultiSurf-GPT delivers more interpretable results by utilizing user input, datasets, and the context-awareness of the papers. For instance, MultiSurf-GPT correctly interprets that "browsing news on his phone" requires the phone to be face-up, aligning with MicroCam's method of using a micro camera on the phone. Although the Multi-Spectrum Camera (SpeCam) offers higher accuracy, its requirement for the phone to be face-down makes it less suitable for this scenario. MultiSurf-GPT seamlessly integrates with the user's current activity, making it practical for everyday use. This demonstrates a thorough understanding of the "activity" context, consistent with Byun et al.'s research [6]. Regarding the statement, "The Tangible Radar method typically requires specialized hardware not commonly found in consumer smartphones," MultiSurf-GPT recognizes that consumer smartphones generally lack radar devices, indicating a solid grasp of the "identity" context. Therefore, MultiSurf-GPT performs better than the original GPT-4o in terms of context awareness.

Table 1: The zero-shot and prompting strategies. <MOD> as a placeholder denotes different modalities (such as radar). <MODEL> as a placeholder denotes different Machine Learning Models (such as SVM, RF). <CLASS> as a placeholder denotes the list of categories of pictures. <PIC1>, <PIC2>...<PICn> as placeholders denote the example for each category. As for zero-shot, the <PIC1>, <PIC2>...<PICn> examples will be deleted.

Data	Task Explaining	Rules
CSV file	The provided CSV is <MOD> data. In the CSV file, columns [0:-1] contain the radar features, and the last column [-1] contains the labels. Build a model (defaulting to using <MODEL>) and return the accuracy. Do not output other text.	[Rules]: Do not output any other text.
Image	The provided picture is <MOD>. Identify the category of this picture from <CLASS> and return only one category (only one category can be returned). For example, <PIC1>, <PIC2>...<PICn> (where n is the number of categories in <CLASS>) are sample images for each category in <CLASS>.	[Rules]: Must return within <CLASS>. Do not output any other text.
Document (paper)	According to the given paper, what equipment did they use, what is the method, and what is the origin usage of the data?	[Rules]: Summarize the method and origin usage each in one complete sentence.

Name	Sensing Method	Task	Recognition Accuracy (%) by MultiSurf-GPT	Usage Domain
Tangible Radar	Objects are placed on or near the Soli sensor, which transmits electromagnetic waves. The sensor captures the reflected waves, revealing the objects' distance, thickness, shape, density, and composition.	Count	68.35	Counting layers of thin objects (e.g., paper, cards, chips).
		Order	73.59	Determining the order of stacked items (e.g., cards).
		Rotation	100.00	Tracking object rotation for control inputs.
		Orientation	90.91	Detecting object orientation (face-up or face-down).
		Identify	90.45	Recognizing objects by their unique radar signal reflection.
		Distance	69.64	Measuring object distance and proximity.
MicroCam	Ensuring the phone's microscope camera is in contact with the surface, and the phone is faced-up. The IMU continuously monitors movement and activates the microscope camera to capture high-resolution images when the phone is stationary.	Object(zero-shot)	20.00	Recognizing objects where the phone is placed (e.g., bed, desk, sofa, cabinet, sink).
		Object(1-shot)	43.33	
		Material(zero-shot)	40.00	Identifying materials of the surfaces (e.g., plush, fabric, leather, wood, ceramic).
		Material(1-shot)	51.11	
SpeCam	The phone face-down flashes seven colors, captures reflected light and sends images to the server for classification.	Material(zero-shot)	5.00	Accurately distinguishing between 30 types of surface materials based on their reflected light properties.
		Material(1-shot)	56.67	

Figure 1: Low-level information on 3 surface sensing methods as recognized and captured by the MultiSurf-GPT framework for high-level context-aware reasoning.

(User Input) Imagine a scenario where a user likes to drink coffee in the morning while sitting at the dining table and browsing news on his phone. At this time, if you need to determine the material of the dining table, which method would be better: Tangible Radar, Micro Camera, or Multi-Spectrum Camera? Please give one or two suggestions.

a

(Original GPT-4o Output)

Micro Camera would be the primary recommendation due to its ease of use, high resolution, and ability to capture fine surface details. If a more detailed material analysis is required, a **Multi-Spectrum Camera** could be considered as an additional method.

b

(MultiSurf-GPT Output)

Micro Camera (MicroCam) is the best method for determining the material of the dining table in the given scenario because:
1. **Convenience:** Allows the phone to remain face-up, enabling the user to continue browsing news without interruption.
2. **Moderate Accuracy:** Provides a reasonable one-shot accuracy for material identification.
3. **Ease of Use:** Seamlessly integrates with the user's current activity, making it practical for everyday use.

Alternative: While the **Multi-Spectrum Camera (SpeCam)** offers higher accuracy, its requirement for the phone to be face-down makes it less suitable for this scenario. The **Tangible Radar** method typically requires specialized hardware that is not commonly found in consumer smartphones.
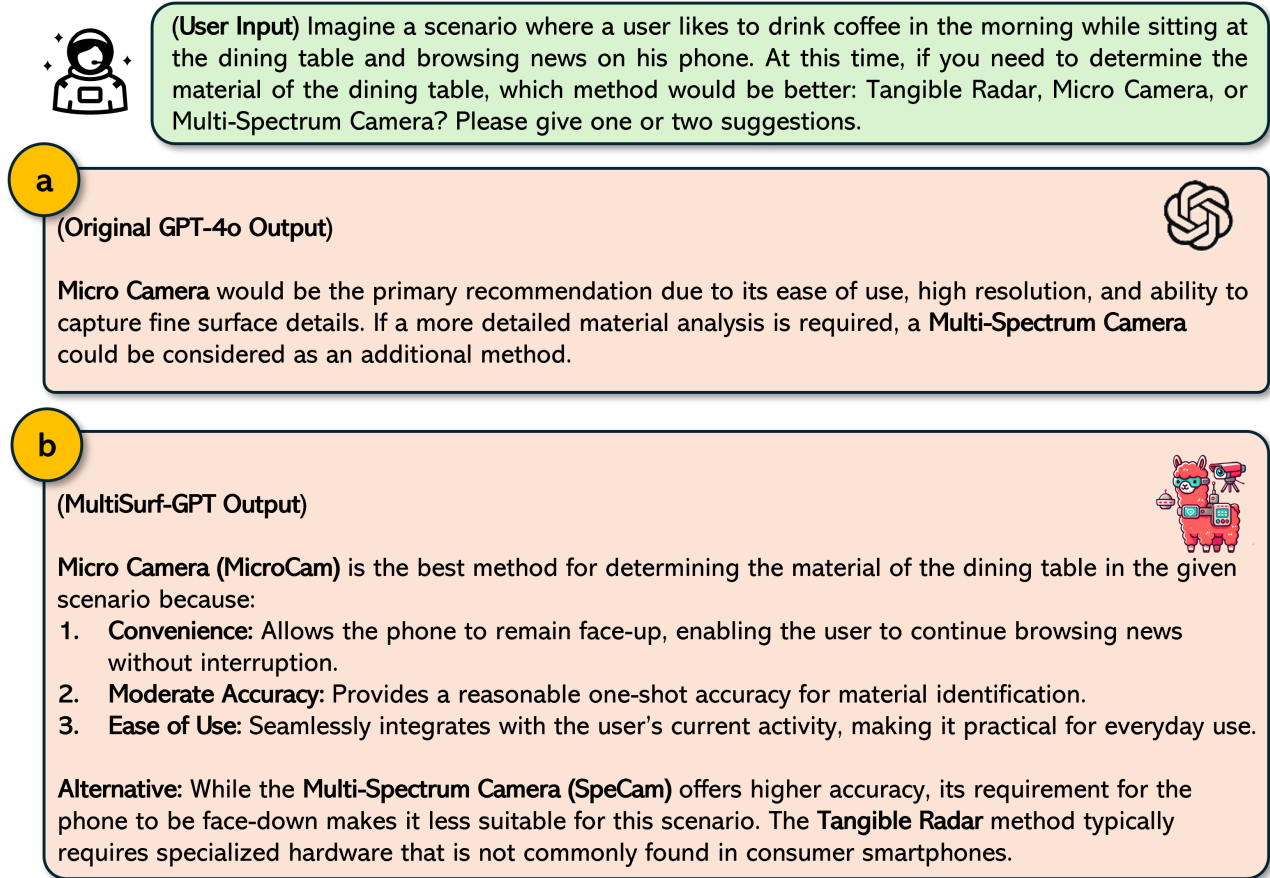
Figure 2: A good experiment case: (a) represents the output by the original GPT-4o model, recommending a Micro Camera as the primary method and a Multi-Spectrum Camera for more detailed analysis if needed. (b) illustrates the output by MultiSurf-GPT, which favors the Micro Camera for its convenience, moderate accuracy, and ease of use, specifically highlighting the practical limitations of the Multi-Spectrum Camera and Tangible Radar for everyday scenarios.

5 CONCLUSION AND FUTURE WORK

We have investigated the integration of multimodal LLMs like GPT-4o through our MultiSurf-GPT framework for processing and interpreting multimodal surface sensing data across radar, microscopy, and multispectral modalities. Utilizing zero-shot and one-shot cueing strategies, MultiSurf-GPT has shown initial success in recognizing low-level information and deriving high-level contextual insights. This deployment of LLMs in mobile contextual awareness paves the way for more sophisticated and integrated applications, offering enhanced user-device interactions through advanced reasoning and data processing capabilities.

In the future, the MultiSurf-GPT framework presents several opportunities for enhancement and further research. Currently, its reliance on cue engineering can limit recognition accuracy, particularly with complex datasets like MicroCam and SpeCam. Future versions could benefit from incorporating instruction fine-tuning to improve recognition accuracy. Additionally, wide-range subjective user experiments would help quantitatively and precisely assess how MultiSurf-GPT contributes to enhanced contextual awareness in surface sensing. These initiatives would not only refine the model's performance but also broaden its practical applications, fostering the development of more adaptive and intelligent mobile applications.

REFERENCES

[1] Gregory D Abowd, Anind K Dey, Peter J Brown, Nigel Davies, Mark Smith, and Pete Steggles. 1999. Towards a better understanding of context and context-awareness. In *Handheld and Ubiquitous Computing: First International Symposium, HUC'99 Karlsruhe, Germany, September 27–29, 1999 Proceedings 1*. Springer, 304–307.

[2] J Augusto, Asier Aztiria, Dean Kramer, and Unai Alegre. 2017. A survey on the evolution of the notion of context-awareness. *Applied Artificial Intelligence* 31, 7-8 (2017), 613–642.

[3] Juan Carlos Augusto. 2022. Contexts and context-awareness revisited from an intelligent environments perspective. *Applied Artificial Intelligence* 36, 1 (2022), 2008644.

[4] Matthias Baldauf, Schahram Dustdar, and Florian Rosenberg. 2007. A survey on context-aware systems. *International Journal of ad Hoc and ubiquitous Computing* 2, 4 (2007), 263–277.

[5] Alberto Berenguer, Adriana Morejón, David Tomás, and Jose-Norberto Mazón. 2024. Leveraging Large Language Models for Sensor Data Retrieval. *Applied Sciences* 14, 6 (2024), 2506.

[6] Hee Eon Byun and Keith Cheverst. 2001. Exploiting user models and context-awareness to support personal daily activities. In *Workshop in UM2001 on User Modeling for Context-Aware Applications*. Citeseer.

[7] Yifei Cheng, Yukang Yan, Xin Yi, Yuanchun Shi, and David Lindlbauer. 2021. Semanticadapt: Optimization-based adaptation of mixed reality layouts leveraging virtual-physical semantic connections. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. 282–297.

[8] Andrea Coletta, Kshama Dwarakanath, Penghang Liu, Svitlana Vyetrenko, and Tucker Balch. 2024. LLM-driven Imitation of Subrational Behavior: Illusion or Reality? *arXiv preprint arXiv:2402.08755* (2024).

[9] Lajos Matyas Csepregi. 2021. The effect of context-aware llm-based npc conversations on player engagement in role-playing video games. *Unpublished manuscript* (2021).

[10] Don Samitha Elvitigala, Yunfan Wang, Yongquan Hu, and Aaron J Quigley. 2023. RadarFoot: Fine-grain Ground Surface Context Awareness for Smart Shoes. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–13.

[11] Hans W Gellersen, Albrecht Schmidt, and Michael Beigl. 2002. Multi-sensor context-awareness in mobile devices and smart artifacts. *Mobile Networks and Applications* 7 (2002), 341–351.

[12] Chris Harrison and Scott E Hudson. 2008. Lightweight material detection for placement-aware mobile computing. In *Proceedings of the 21st annual ACM symposium on User interface software and technology*. 279–282.

[13] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their applications* 13, 4 (1998), 18–28.

[14] Yongquan Hu, Zhaocheng Xiang, Lihang Pan, Xiaozhu Hu, Yinshuai Zhang, and Aaron J Quigley. 2023. Exploring the Adaptation of Mobile GUI to Human Motion Status. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*. 145–147.

[15] Yongquan Hu, Hui-Shyong Yeo, Mingyue Yuan, Haoran Fan, Don Samitha Elvitigala, Wen Hu, and Aaron Quigley. 2023. Microcam: Leveraging smartphone microscope camera for context-aware contact surface sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (2023), 1–28.

[16] Yongquan Hu, Mingyue Yuan, Kaiqi Xian, Don Samitha Elvitigala, and Aaron Quigley. 2023. Exploring the design space of employing ai-generated content for augmented reality display. *arXiv preprint arXiv:2303.16593* (2023).

[17] Guang-Li Huang, Arkady Zaslavsky, Seng W Loke, Amin Abkenar, Alexey Medvedev, and Alireza Hassani. 2022. Context-aware machine learning for intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems* 24, 1 (2022), 17–36.

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

[19] Xiang Li, Congcong Wen, Yuan Hu, Zhenghang Yuan, and Xiao Xiang Zhu. 2024. Vision-language models in remote sensing: Current progress and future trends. *IEEE Geoscience and Remote Sensing Magazine* (2024).

[20] M Jehanzeb Mirza, Leonid Karlinsky, Wei Lin, Sivan Doveh, Jakub Micorek, Mateusz Kozinski, Hilde Kuhene, and Horst Possegger. 2024. Meta-Prompting for Automating Zero-shot Visual Recognition with LLMs. *arXiv preprint arXiv:2403.11755* (2024).

[21] Xiaomin Ouyang and Mani Srivastava. 2024. LLMSense: Harnessing LLMs for High-level Reasoning Over Spatiotemporal Sensor Traces. *arXiv preprint arXiv:2403.19857* (2024).

[22] Daniel Salber. 2000. Context-awareness and multimodality. In *Colloque sur la multimodalité*.

[23] Maximilian Schrapel, Philipp Etgeton, and Michael Rohs. 2021. SpectroPhone: Enabling Material Surface Sensing with Rear Camera and Flashlight LEDs. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–5.

[24] Daniel P Siewiorek, Asim Smailagic, Junichi Furukawa, Andreas Krause, Neema Moraveji, Kathryn Reiger, Jeremy Shaffer, and Fei Lung Wong. 2003. SenSay: A Context-Aware Mobile Phone.. In *ISWC*, Vol. 3. Citeseer, 248.

[25] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[26] Mark Weiser. 1991. The Computer for the 21 st Century. *Scientific american* 265, 3 (1991), 94–105.

[27] Xuhai Xu, Bingshen Yao, Yuanzhe Dong, Hong Yu, James Hendler, Anind K Dey, and Dakuo Wang. 2023. Leveraging large language models for mental health prediction via online text data. *arXiv preprint arXiv:2307.14385* (2023).

[28] Xing-Dong Yang, Tovi Grossman, Daniel Wigdor, and George Fitzmaurice. 2012. Magic finger: always-available input through finger instrumentation. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. 147–156.

[29] Hui-Shyong Yeo, Gergely Flamich, Patrick Schrempf, David Harris-Birtill, and Aaron Quigley. 2016. Radarcat: Radar categorization for input & interaction. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 833–841.

[30] Hui-Shyong Yeo, Juyoung Lee, Andrea Bianchi, David Harris-Birtill, and Aaron Quigley. 2017. Specam: Sensing surface color and material with the front-facing camera of a mobile device. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–9.

[31] Hui-Shyong Yeo, Ryosuke Minami, Kirill Rodriguez, George Shaker, and Aaron Quigley. 2018. Exploring tangible interactions with radar sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–25.

[32] Jiadi Yu, JIaming Zhao, Yingying Chen, and Jie Yang. 2015. Sensing ambient light for user experience-oriented color scheme adaptation on smartphone displays. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*. 309–321.

[33] Tongxin Yuan, Zhiwei He, Lingzhong Dong, Yiming Wang, Ruijie Zhao, Tian Xia, Lizhen Xu, Binglin Zhou, Fangqi Li, Zhuosheng Zhang, et al. 2024. R-Judge: Benchmarking Safety Risk Awareness for LLM Agents. *arXiv preprint arXiv:2401.10019* (2024).

[34] Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. 2024. Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain. *arXiv preprint arXiv:2401.16822* (2024).

[35] Yang Zhang, Chouchang Yang, Scott E Hudson, Chris Harrison, and Alanson Sample. 2018. Wall++ room-scale interactive and context-aware sensing. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–15.