# Sensor2Text: Enabling Natural Language Interactions for Daily Activity Tracking Using Wearable Sensors

WENQIANG CHEN*, Chinese Academy of Sciences, SIAT Massachusetts Institute of Technology, USA
JIAXUAN CHENG*, Massachusetts Institute of Technology, USA
LEYAO WANG, Vanderbilt University, USA
WEI ZHAO, University of SIAT, Shenzhen, China
WOJCIECH MATUSIK, Massachusetts Institute of Technology, USA

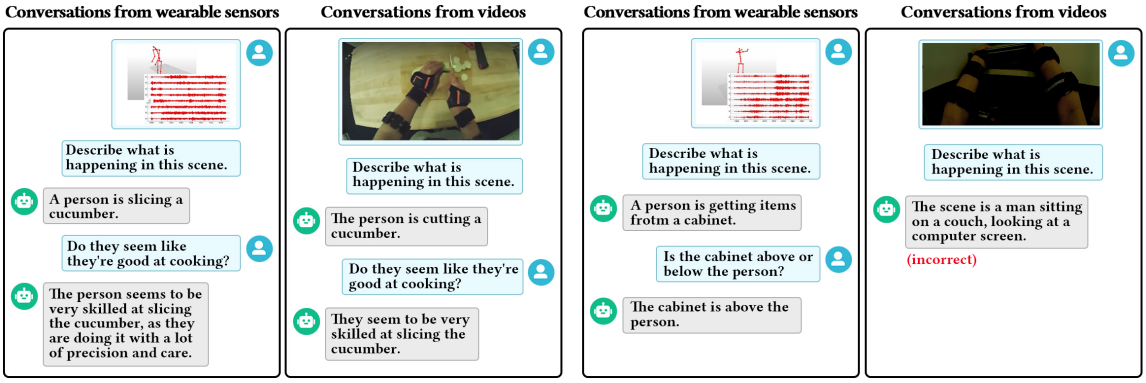



Fig. 1. Sensor2Text demonstrates an understanding of wearable sensor inputs and interactive capability, compared with conversations generated from existing vision models on the ground truth videos. Sensor2Text accurately comprehends daily activities through wearable sensor data, circumventing video-related constraints like inadequate lighting or obstructed viewpoints.

Visual Question-Answering, a technology that generates textual responses from an image and natural language question, has progressed significantly. Notably, it can aid in tracking and inquiring about daily activities, crucial in healthcare monitoring, especially for elderly patients or those with memory disabilities. However, video poses privacy concerns and has a limited field of view. This paper presents Sensor2Text, a model proficient in tracking daily activities and engaging in conversations using wearable sensors. The approach outlined here tackles several challenges, including low information density in wearable sensor data, insufficiency of single wearable sensors in human activities recognition, and model's limited capacity for Question-Answering and interactive conversations. To resolve these obstacles, transfer learning and student-teacher networks are utilized to leverage knowledge from visual-language models. Additionally, an encoder-decoder neural network model is devised to jointly process language and sensor data for conversational purposes. Furthermore, Large Language Models are also utilized to enable interactive capabilities. The model showcases the ability to identify human activities and engage in

* The first and second authors contributed equally to this work.

Authors' addresses: Wenqiang Chen*, Chinese Academy of Sciences, SIAT and Massachusetts Institute of Technology, USA, wenqiang@mit.edu; Jiaxuan Cheng*, Massachusetts Institute of Technology, USA, jasonc75@mit.edu; Leyao Wang, Vanderbilt University, USA, leyaowang8@gmail.com; Wei Zhao, University of SIAT, Shenzhen, China, zhao.wei@siat.ac.cn; Wojciech Matusik, Massachusetts Institute of Technology, USA, wojciech@csail.mit.edu.


2474-9567/2024/12-ART192
https://doi.org/10.1145/3699747

Q&A dialogues using various wearable sensor modalities. It performs comparably to or better than existing visual-language models in both captioning and conversational tasks. To our knowledge, this represents the first model capable of conversing about wearable sensor data, offering an innovative approach to daily activity tracking that addresses privacy and field-of-view limitations associated with current vision-based solutions.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

Additional Key Words and Phrases: human activity recognition, wearable sensors, large language models

## 1 INTRODUCTION

Visual Question Answering is an exciting task in computer vision and natural language processing that has gained much progress in recent years; it involves generating a text response given an image and natural language question [4]. Notably, visual question-answering can be applied in tracking personal daily activities, which shows great promise for healthcare monitoring, especially for elderly people or patients with Alzheimer's. While traditional monitoring systems can only send raw footage generated by monitoring cameras to caregivers, current question-answering systems enable conversational interactions with the users. This capability allows caregivers to inquire about various aspects of the target's well-being, such as medication adherence, dietary habits, sleep patterns, and more.

However, using video as the input modality to question-answering has several issues. Video suffers from obstructions, restricted viewpoints, and vulnerability to inadequate lighting, all of which are common occurrences in daily activities monitoring. More importantly, using constant video surveillance has privacy concerns and may make the subject feel uncomfortable. Therefore, it is crucial to devise an innovative approach to replace the reliance on cameras, which enables the collection of data on subjects' daily activities while also preserving their privacy.

It is observed that sensor-based data, particularly those obtained from wearable sensors, can be a promising alternative. Wearable sensors offer several advantages over video surveillance. Firstly, they allow the subject to roam freely outside of designated areas, as opposed to video which requires the subjects to remain within the camera coverage zones. Additionally, compared with video cameras, wearable sensors can provide more information regarding one's health and physical activity, making it a better option for healthcare monitoring. Finally, they are less invasive in privacy compared to video cameras.

In this paper, we propose Sensor2Text, a sensor language model capable of conditioning itself on wearable sensor data to perform the aforementioned question-answering, illustrated in Figure 1. Sensor2Text is interactive and versatile for subjects and caregivers to track daily activities. Additionally, it functions without requiring any visual input, making it less obtrusive than video monitoring and applicable in dark or occluded settings. Furthermore, compared with existing video and sensor captioning models [30, 53, 73], Sensor2Text stands out for its chat capabilities and proficiency in question-answering tasks, enhancing its utility in tracking daily activities.

However, developing a sensor-language model like Sensor2Text presents several challenges. Firstly, wearable sensors lack the density of information found in video data regarding human activities. To overcome this, we employ cross-modality supervision using teacher-student networks. In this approach, a model trained on visual camera data acts as a teacher, efficiently transferring its knowledge to models analyzing sensor data after relatively few sensor examples.

In addition, a single wearable sensor is often insufficient when differentiating between similar activities. To fix this, we use multiple wearable devices to track human activities. We jointly process the data modalities using a multi-modal encoder-decoder neural network architecture and align the temporal dynamics between the modalities with time-dependent output tokens.

Furthermore, even if the model can comprehend sensor data, it lacks the capacity for Question-Answering or interactive conversations. To tackle this obstacle, Large Language Models are integrated into Sensor2Text to endow it with conversational abilities.

We conducted a comprehensive evaluation of the Sensor2Text model, assessing its performance through both qualitative and quantitative analyses. The results demonstrate that Sensor2Text exhibits robust capabilities in responding accurately to queries and providing detailed descriptions based on wearable sensor data input. In our tests, Sensor2Text achieved captioning scores comparable to those of leading video-language models, with an average BLEU score of 73.0 and a CIDEr score of 62.2. This closely aligns with the scores of existing vision-language models, which average a BLEU score of 76.7 and a CIDEr score of 62.5, underscoring its competitive sensor recognition capabilities. Moreover, Sensor2Text shows remarkable performance in challenging conditions, such as low-light or occluded environments where vision-based models are ineffective. Furthermore, it demonstrates strong generalization across different users and effective utilization of multiple sensor data modalities. Additionally, ablation studies were conducted to ascertain the contributions of individual components to the model's overall efficacy and training methodology.

In summary, the major contributions of our work are:

- To the best of our knowledge, we are the first to create a chatbot capable of engaging in Q&A and conversations regarding various human daily activities, achieved through comprehension of multi-modal wearable sensor data.
- We design a transformer-based neural network encoder, including late fusion and a 1D convolutional tokenizer, and use auxiliary training losses to extract useful sensor representations while efficiently transferring knowledge from video.
- We evaluate our model through extensive experiments, including qualitative and quantitative assessments, generalization tests on unseen subjects, and ablation studies.

## 2 RELATED WORK

Our study is closely connected to research within the domains of Human Activity Recognition, Video Captioning, and Multi-modal Large Language Models. Below, we provide an in-depth exploration of these areas, elucidating how Sensor2Text leverages insights from prior research to pioneer a novel approach enabling conversations and question-answering about content in wearable sensors. This innovative approach mitigates privacy concerns and viewpoint limitations in existing vision-based solutions, making it a promising alternative in applications such as healthcare monitoring.

### 2.1 Human Activity Recognition

The domain of Human Activity Recognition (HAR)[19], which focuses on identifying human actions through sensor data analysis, has witnessed notable progress in recent years, particularly in leveraging machine learning techniques for interpreting human activities from sensor data. Researchers have utlized machine learning techniques such as kernel discriminant analysis and Long Short-Term Memory (LSTM)-based Neural Structured Learning [35, 65], Decision Trees, Support Vector Machines, and Neural Networks [61] to recognize activities from smartphone data. Another area of interest involves the use of radio signals interacting with human bodies to comprehend movements and behaviors. Specifically, previous studies have demonstrated the use of radio signals for tasks such as location tracking [1, 30, 48], extraction of 3D skeletons of nearby individuals [78], and action classification [47]. Finally, some studies also employ wearable sensors and deep learning techniques for HAR and other related tasks [7, 11–18, 20, 22–25, 28, 34, 39, 55, 62, 70**?** ], including unsupervised domain adaptation [9], contrastive predictive coding [38], and deep unsupervised clustering [54].

While recent studies have made significant progress in Human Activity Recognition (HAR), the predominant focus lies in classifying recognized activities. However, Sensor2Text is capable of performing captioning and utilizing information from multiple modalities of wearable sensor data. Additionally, our approach endeavors to perform question-answering and have conversations regarding the content gleaned from wearable sensor data during daily activities rather than only simple text messages from captioning.

## 2.2 Video Captioning

Video captioning refers to the generation of descriptive text for video content. One strategy for video captioning involves applying image captioning models to uniformly sampled frames [32, 49]. Recent advancements employ recurrent neural networks (RNNs) [68], attention mechanisms [66], hierarchical architectures [69, 75] to capture more temporal intricacies in the video. Furthermore, recent work studies the integration of audio and video comprehension through advancements in multi-modal learning [10, 73].

Another development pertinent to video captioning is Vision-Language Pre-training (VLP), which aims to enhance the performance of multi-modal foundation models across various vision-language tasks through unsupervised learning. For instance, CLIP [60] utilizes conservative loss between image/text pairs to pre-train a model on internet-scraped data. These models often adopt architectures like dual-encoders [56, 60, 72], fusion-encoders [63], or a hybrid of both as encoder-decoder architectures [46]. VLP models can be applied across a range of vision-based tasks, including classification and captioning, among others.

While great advancements have been made in video captioning, our work, in contrast, is intended to employ wearable sensor data for conversational interactions with the user. However, the sensor-language model continues to leverage advancements in video captioning and VLP by employing visual language models in transfer learning to distill the learned knowledge to wearable sensor data. After training, the user will be able to conduct Q&A conversations using wearable sensor data as input, without requiring any visual input.

## 2.3 Multi-modal Large Language Models

Large Language Models (LLMs) have garnered significant attention in recent years for their ability to comprehend and generate human-like text. There is a growing interest in extending their capabilities to process modalities of inputs other than language. One approach is to use LLMs as controllers, with existing multi-modal models serving as tools [8, 40, 41, 57, 74]. In this setup, the LLM decides which tools to call based on the user's text instruction and incorporates results from these multi-modal models. Other work focuses on developing fundamental multi-modal LLMs by aligning the representation space of pre-trained unimodal models with that of LLMs, allowing the language model to natively understand other data modalities rather than through a proxy. Models like BLIP2 [46] and LLaVA [51] exemplify this approach.

Despite these advancements, the majority of multi-modal LLMs primarily accept images as input [76]. Some studies have explored processing other data modalities, such as 3D point clouds [36] or audio [77], but few have attempted to extend the capabilities of LLMs to comprehend wearable sensor data. Our proposed model, Sensor2Text, aims to bridge this gap by enabling the understanding of wearable sensor data and its integrate them withLLMs.

In summary, our research extends the existing body of literature on Human Activity Recognition, Video Captioning, Vision-Language Pre-training (VLP), and multi-modal Large Language Models. What sets our model apart is its utilization of wearable sensor data, rather than conventional visual data, to perform Q&A and chat about an individual's daily activities.

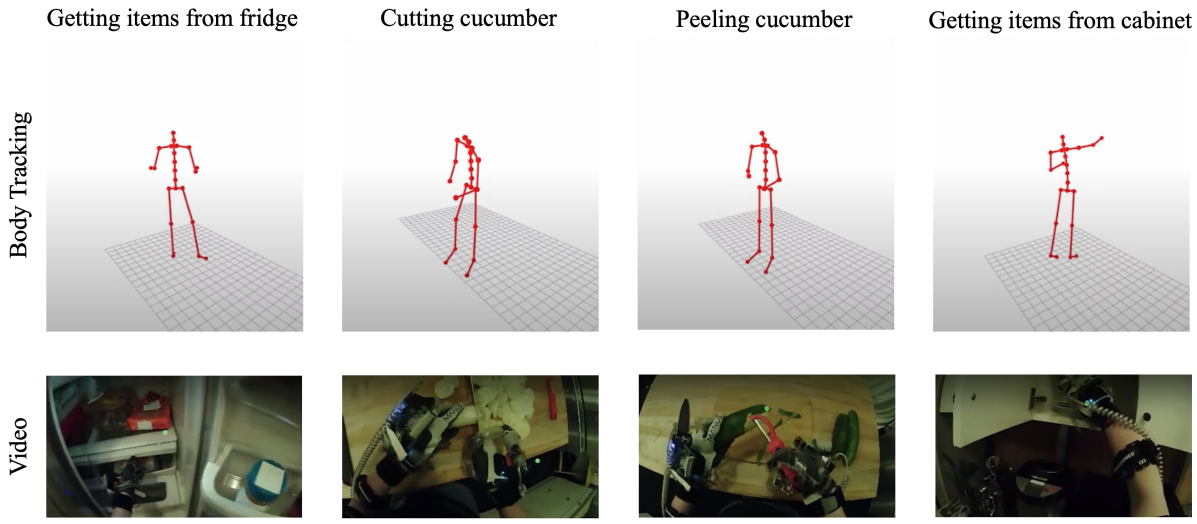



Fig. 2. Visualization of body tracking data using wearable sensors. Different activities display distinct body positions, demonstrating the feasibility of using wearable sensors to identify daily activities.

## 3 PRELIMINARY STUDY

This section first explores the capability of wearable sensor data in recognizing and distinguishing daily activities. It then assesses the feasibility of training such a sensor model due to the scarcity of comprehensive multi-modal wearable sensor datasets.

### 3.1 Sensor Characteristics for Activity Tracking

We observe that wearable sensor data can effectively be used to identify human activities for the following reasons. Firstly, activities that may appear similar through video cameras often yield distinct sensor readings. Illustrated in Figure 2, while actions like peeling or cutting cucumbers may seem alike in video images, the corresponding human skeleton representations derived from sensor data exhibit notable differences.

Additionally, wearable sensors reveal dynamic variations. Each wearable sensor gathers time series data rather than static snapshots, capturing a subject's movement patterns over time that can further aid in distinguishing similar activities.

Furthermore, integrating multiple modalities from wearable sensors enhances the differentiation of human activities. While certain activities may exhibit similar body tracking readings, they may yield entirely different readings from other wearable sensors. For instance, although body tracking data excels in distinguishing activities with various movements, it may struggle to discern certain nuances, such as differentiating between cutting a cucumber and a potato. However, by integrating muscle activation sensors that monitor electromyography signals, it becomes possible to differentiate between the objects the subject is holding [21, 31], effectively distinguishing the action of cutting a potato from cutting a cucumber. By processing multiple modalities of sensor data through a multi-modal deep learning architecture, it becomes possible to make use of each modality in determining the details of the activity.

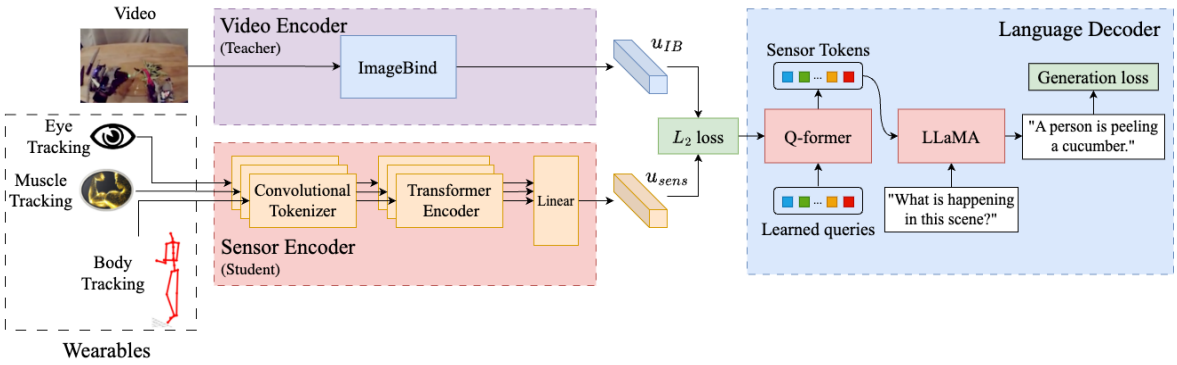



Fig. 3. Architecture diagram of Sensor2Text.

To summarize, through the integration of diverse and carefully selected data modalities alongside appropriate temporal-aware processing, it is possible that sensor data can be utilized to classify daily activities with performance equal to, or even surpassing, that achieved with video data.

## 3.2 Feasibility

A significant concern regarding the feasibility of constructing a sensor-language model is the scarcity of large-scale wearable sensor datasets. In contrast, visual-language models benefit from abundant datasets, facilitating extensive training or fine-tuning for robust performance in captioning tasks. Consequently, an insightful approach emerges: we can transfer the knowledge acquired from training visual-language models to sensor data through transfer learning and teacher-student networks.

Indeed, prior research [36, 37] demonstrates the feasibility of leveraging large visual-language datasets for downstream fine-tuning by aligning the encoder of new modality data with an existing visual encoder. This significantly mitigates the need for an extensive amount of wearable sensor data. Accordingly, employing wearable sensor data to train a sensor language model is feasible.

# 4 METHOD

Sensor2Text is a novel model designed to generate textual responses based on sensor data and textual prompts. The primary challenges in developing such a model are the low information density of single modal sensor data and the limited sizes of existing sensor datasets. To address these challenges, a unique architecture is proposed, illustrated in Figure 3. The model consists of two main subcomponents: a sensor encoder and a language decoder. The sensor encoder takes multiple modalities of sensor data as input and outputs a learned representation vector that incorporates information from all sensors. The language decoder then takes this representation along with the language prompt and generates the textual response.

To efficiently train the model despite the lack of sensor data, a vision encoder is introduced as a teacher network to train the sensor encoder. Additionally, transfer learning techniques are employed to train the language decoder. To facilitate the training process, a dataset is carefully selected to include videos, labels, and multiple sensor data modalities.

The following subsections provide a detailed description of the architecture, including the sensor encoder and language decoder, as well as the procedure used to train both components.

### 4.1 Sensor Encoder

This section introduces the sensor encoder, which extracts useful features from time series data collected by the various wearable sensors and integrates them into a single representation.

We utilize distinct encoders to extract features from each sensor data modality separately before combining them with late fusion. A transformer encoder [66] is applied to each data modality due to its excellent performance in capturing intricate dependencies within the sequential data. Each sequence of wearable sensor data, denoted as $x^{(i)} \in \mathbb{R}^{T_i \times d_i}$ for the $i$th sensor modality, where $d_i$ denotes the number of features and $T_i$ represents the number of timesteps, is processed via a 1D convolutional tokenizer. This converts the raw sensor data into tokens that are more easily ingested by the downstream transformer. The sequential data is segmented into fixed-length windows and a learned 1D convolution is applied to each window. The resulting token for each window is represented as $u_j^{(i)} = \theta_{\text{tok}}^{\top} x_{s*j,\ldots,s*j+d-1}^{(i)}$, where $d$ indicates the window size, $s$ denotes the stride between consecutive windows, $\theta_{\text{tok}} \in \mathbb{R}^{d \times d_{\text{encoder}}}$ represents the learnable weights of the tokenizer, and $u_j^{(i)} \in \mathbb{R}^{d_{\text{encoder}}}$ represents the output tokens.

The tokenizer then applies a sine/cosine positional encoding onto each token to inject temporal information into each token, rendering different positions distinguishable to the downstream transformer. The positional encoding for the $j^{th}$ position and $d^{th}$ dimension is given as $\text{PE}(j, d)$, where:

$$\text{PE}(j, 2k) = \sin(j/10000^{2k/d_{\text{encoder}}})$$
$$\text{PE}(j, 2k+1) = \cos(j/10000^{2k/d_{\text{encoder}}})$$

Each positional encoding is directly added onto the token sequence for each input data modality, resulting in a modified sequence of position-aware tokens denoted as $u'_j = u_j + \text{PE}(j)$. A learnable classification token, denoted as `[CLS]`, is prepended to this sequence. After processing in the downstream sequence-to-sequence transformer, the output of this token will serve as the single representation for the entire sequence. Consequently, we obtain a sequence of tokens $\mathbf{u} = ([\text{CLS}], u'_1, \ldots, u'_n)$.

Next, the position-aware tokens are processed in a transformer encoder to obtain an intermediate representation for each sensor. The transformer encoder consists of $N$ layers of alternating self-attention and feedforward layers, effectively capturing both long and short-term dependencies in the sensor tokens. The output vector of the `[CLS]` token is extracted to obtain an intermediate representation for each sensor modality, $y^{(i)} = \text{Transformer}(\mathbf{u}^{(i)})$, where $y^{(i)} \in \mathbb{R}^{d_{\text{encoder}}}$.

Finally, the late fusion technique is employed to obtain the ultimate representation vector. This process involves concatenating the representation vectors of each sensor and then applying a final feedforward layer to capture potential interdependencies among all sensor modalities. The outcome is a single representation vector, $y_{\text{sens}} \in \mathbb{R}^{d_{\text{output}}}$, which encapsulates all relevant information from the wearable sensor data, where $d_{\text{output}}$ is the dimension of the final representation vector.

### 4.2 Auxiliary Training Objective

We leverage a visual encoder as a teacher network to train the sensor encoder. Through this method, the sensor encoder can effectively distill valuable information from the visual encoder while avoiding the instability, high computational cost, and susceptibility to overfitting that comes from training the encoder with an end-to-end text generation loss. For this purpose, the ImageBind model [33] is utilized as the designated teacher network. ImageBind has the capacity to encode diverse modalities such as visual, audio, IMU, thermal, depth, and text data into a unified embedding space. Its proficiency in learning representations that transcend modality-specific boundaries makes it an ideal candidate to act as a teacher network.

Specifically, the number of output features of the sensor encoder, $d_{\text{encoder}}$, is set equal to the output dimension of ImageBind and train the sensor encoder to align with the visual encoder using $L_2$ loss on paired sensor/video examples. Adhering to the methodology adopted in ImageBind for video processing, two representative frames are selected for each video for processing. For video $i$, denoted as $V_i = [v_0; ...; v_{n_i-1}]$, where $n_i$ is the number of frames, we select $v_0$ and $v_{n_i/2}$ to be inputted into ImageBind. The total loss is formulated as follows:

$$\mathcal{L}_{\text{aux}}(x;\theta) = \sum_{(x_i,V_i)\in\mathcal{D}} \left\|\text{Encoder}(x_i;\theta) - \text{ImageBind}(V_{i,0}, V_{i,n_i/2})\right\|^2 + R(\theta)$$

Here, $x_i$ and $V_i$ represent paired sensor and video data from the dataset $\mathcal{D}$, with $\theta$ denoting the encoder's weights and R denoting a regularization term. Through this approach, the sensor encoder can generate meaningful representations without necessitating the execution of the entire Sensor2Text model in an end-to-end manner. After this training phase, the weights of the sensor encoder are frozen for the duration of the remaining training process.

## 4.3 Language Decoder

After acquiring the feature representation for a segment of sensor data, it becomes essential to devise a method to process both the input textual prompt and the sensor data to generate a response. To accomplish this, the LLaMA Large Language Model is employed as the language decoder [64]. Furthermore, a bridging layer is incorporated to reconcile our sensor representation with the subsequent language decoder.

*4.3.1 Primer on Querying Transformer.* Prior to processing the encoded sensor representation with LLaMA, we utilize the Querying Transformer [46] to bridge the sensor encoder's outputs to LLaMA's embedding space. The Querying Transformer (Q-former) is a bottleneck architecture designed to distill textual information from the outputs of an upstream encoder. The Q-former is a transformer-based model that utilizes a set of learned queries as input and uses cross-attention to interact with the embeddings of the upstream encoder. Compared to traditional methods, such as employing a linear layer or feedforward network, the Q-former is able to generate any number of tokens for the downstream language decoder. Moreover, the use of attention and cross-attention allows the Q-former to learn and output complex interdependencies between the output tokens. As a result, the tokens produced are semantically coherent to the downstream LLM.

We initialize the Q-former with pre-trained weights to leverage the existing vision-language knowledge inherited from pre-training tasks [46] and further end-to-end text generation finetuning. Specifically, the pre-trained Q-former from VideoLLaMA's [77] omni-modal (AL) branch is utilized because it has undergone the aforementioned pre-training as well as further finetuning on vision-language tasks using ImageBind as the visual encoder, serving as the ideal base model to transfer knowledge to sensor data.

*4.3.2 Language Decoder.* The language decoder not only interprets data from the sensor encoder but also handles conversational tasks involving reasoning, leveraging world knowledge, and following instructions. Hence, we opt for a pre-trained Large Language Model (LLaMA-7B) [64] as our optimal choice. LLaMA-7B, the most lightweight model in the LLaMA family, is well-suited for our computationally intensive training process.

Similar to other transformer-based large language models [3, 58], LLaMA employs a text tokenizer to generate embeddings from an input text sequence, represented as $\text{Tokenizer}(x) \in \mathbb{R}^{T\times d_{\text{LLaMA}}}$, before processing it in the transformer layers. Here, $T$ is the number of tokens in the text sequence and $d_{\text{LLaMA}}$ indicates the size of each token embedding. The output of the language model can then be written as $y = \text{LLaMA}(\text{Tokenizer}(x))$.

To incorporate the sensor data, we apply the Querying Transformer to the output of the sensor encoder to extract sensor embeddings and prepend them to the text embeddings. The resulting output is $y = LLaMA(QFormer(\mathbf{u}_{sens})\oplus Tokenizer(x))$. Here, $\mathbf{u}_{sens}$ is the output from the sensor encoder, and $QFormer(\mathbf{u}_{sens}) \in \mathbb{R}^{K\times d_{\text{LLaMA}}}$ is the output

from the Q-former, where $K$ denotes the number of learned queries in the Q-former and $d_{\text{LLaMA}}$ denotes the size of LLaMA's embedding space. Together, the computed sensor tokens and input text tokens allow the model to respond conditioned on sensor inputs.

*4.3.3 Temporal-Aware Tokens.* The sensor representation encapsulates information regarding events within a sequence of sensor data but lacks temporal clarity for the LLM. However, comprehending daily activities necessitates differentiating the chronological sequence of events, which may be further inquired by the user in detail. To tackle this issue, $n$ segments of equal length are sampled from a specified time interval of wearable sensor data. Each segment is then encoded to yield $\mathbf{u}_{sens,t}$ for $t \in [1, ..., n]$. These encoded segments are concatenated with the output tokens from all $n$ intervals when inputted into LLaMA. The final output is represented as $y = LLaMA(QFormer(\mathbf{u}_{sens,1}) \oplus ... \oplus QFormer(\mathbf{u}_{sens,n}) \oplus Embedding(x))$. Given that each of the $n$ sequences of tokens carries meaningful information, different values for $n$ can be employed throughout training and inference. This adaptability enables Sensor2Text to handle variable lengths of wearable sensor data sequences.

Finally, the outputs of all of the $QFormer(\mathbf{u}_{sens,i})$ are encapsulated in natural language, enabling us to leverage more of LLaMA's inherent capabilities. An illustrative prompt is shown in Figure 4.

| Stage | Prompt / Ground Truth |
|---|---|
| Cross-Modal Training | **LLM Input:**<br>[INST] <<SYS>> You are a helpful language and vision assistant. You are able to understand the visual content that the user provides, and assist the user with a variety of tasks using natural language. <</SYS>><br>Open your eyes and imagine you see: <sensor embeddings>. Provide a brief description of the scene. [/INST]<br>**Ground Truth:**<br>A person is slicing a cucumber. |
| Instruct Finetuning | **LLM Input:**<br>[INST] <<SYS>> You are a helpful language and vision assistant. You are able to understand the visual content that the user provides, and assist the user with a variety of tasks using natural language. <</SYS>><br><image embeddings> What objects are the main focus of the image? [/INST]<br>**Ground Truth:**<br>The main focus of the image is a collection of stuffed toy bears. |

Fig. 4. Example prompts and ground truths for the two aforementioned finetuning stages. After the text is tokenized into textual embeddings, the image or sensor embeddings outputted by the Querying Transformer are inserted at the indicated location, allowing the language model to process both the text and the image or sensor information when producing a response. For stage 1, using our hyperparameters, there are 64 sensor embeddings and approximately 100 text embeddings (depending on the training prompt). The same system prompt is used as the pre-trained vision model to maximize the effectiveness of transfer learning.

*4.3.4 Cross-Modal Training.* The language decoder is trained by employing a text generation loss over sensor/caption pairs. This loss is computed utilizing the logits generated by LLaMA for each token in its vocabulary and calculating a cross-entropy loss against the ground truth token. The total loss across a sequence of input tokens $\{x_i\}_{i=1,...,n}$ and output tokens $\{y_i\}_{i=1,...,m}$ is the cumulative sum of cross-entropy losses for each output token, conditioned on the preceding sequence of tokens. These tokens comprise a prompt, as illustrated in Figure

4, and the outputs of the Q-former conditioned on wearable sensor data. The total loss is expressed as follows:

$$\mathcal{L}_{\text{gen}}(x, y; \theta) = -\sum_{i=1}^{m} \log p_\theta(y_i | x, y_1, ..., y_{i-1})$$

Here, $\theta$ denotes the weights of the language decoder and $p_\theta$ is the probability that LLaMA assigns a given token. The loss is efficiently computed using a causal attention mask. This mechanism ensures that the model calculates output logits for a token-based solely on preceding tokens, allowing simultaneous computation of all summands in the generation loss for a given input/output sequence.

During training, LLaMA is frozen to preserve its language modeling abilities. In addition, the sensor encoder is frozen to promote training stability and uphold alignment between the embedding space of the sensor encoder and ImageBind's visual encoder. The loss incurred in text generation is transmitted through LLaMA to the Q-former via the tokens it generates in the input token sequence $\{x_n\}$.

Lastly, we rephrase activity labels into full, natural-language captions. This approach effectively leverages LLaMA's existing textual reasoning abilities, avoiding a substantial decline in language performance that naive fine-tuning might cause. The Q-former is trained using text generation loss with the modified input/output pairs and subsequently learns to produce tokens meaningful to LLaMA and prompt accurate generation outputs after the training.

*4.3.5 Noise Injection for Robustness.* Visual LLMs trained in a similar fashion [46, 51, 77] utilize much larger datasets than existing wearable sensor datasets. Consequently, the language decoder may be less robust, susceptible to overfitting, and exhibit significantly different behaviors with minor variations in the input distribution. To mitigate this issue, noise injection is incorporated into the outputs of the Q-former before feeding them into LLaMA. Noise injection involves adding a random Gaussian vector with a small variance to the inputs, enhancing the model's resilience to minor input alterations. Specifically, we implement

$$QFormer'(\mathbf{u}_{sens})_i = QFormer(\mathbf{u}_{sens})_i + \mathbf{X}_i, \mathbf{X}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$$

where $QFormer(\mathbf{u}_{sens})_i$ denotes the i-th token generated by the Q-former, $\sigma^2$ represents a small variance, and $I_n$ is an identity matrix with n rows and n columns. Since adjacent points in LLaMA's embedding space share similar semantic meanings, this approach ensures robustness without substantially altering LLaMA's perception of its inputs in the training stage. Following the incorporation of noise injection, our model demonstrates improved performance on unseen data, as detailed in subsequent sections.

*4.3.6 Instruct Finetuning.* After completing the initial stage of cross-modal training, the model's performance in conversational contexts, such as responding to user inquiries, diminishes, as it has been trained solely to output activity labels. To address this issue, a second stage of fine-tuning is introduced to restore instruction-following capabilities. In this phase, the model is finetuned using question-answer pairs and example conversations. However, due to the lack of such a dataset with wearable sensor data, a visual instruction dataset is utilized instead. Since the sensor encoder has been trained with the ImageBind visual encoder as a teacher, the language decoder can be finetuned using ImageBind as the encoder on a vision instruct dataset. The Q-former is fine-tuned using text generation loss on image/conversation pairs, using ImageBind to extract visual embeddings. Through this process, the model retains its understanding of sensor data while also regaining instruction-following capabilities. After training is completed, the model can engage in conversations about sensor data without prior exposure to sensor-based discussions.

## 4.4 Training Procedure and Implementation Details

We select the datasets ActionSense [29] and MMAct [43] to evaluate our model. Each consists of subjects performing various daily activities while collecting wearable sensor data, corresponding video footage, and

textual activity labels; further details regarding each dataset are elucidated in the following section. We sample the sensor data using a constant sampling rate of 50Hz or 100Hz, depending on the sensor data modality. We further apply preprocessing steps, such as filtering and normalization, to each wearable sensor data modality, as specified in the following section. Each video is cropped to 224 by 224 pixels and values across each color channel are normalized. These steps ensure that negative effects from faulty sensor readings or video footage are minimized.

We select paired sensor and video clips with a length of 2 seconds to train the sensor encoder. We set the size and stride of each 1D convolutional tokenizer in the model to 10. The sensor encoders for each modality consist of 6 transformer encoder layers, each with a hidden size of 512. Additionally, they feature a feedforward network with a hidden layer size of 2048 and multi-head attention with 8 attention heads and a dimension of 64 per head. Layer Normalization [5] is applied after each attention and feedforward layer in the transformers. Dropout regularization with a dropout rate of 0.1 is also employed. The encoder is trained using the Adam optimizer [42] with a learning rate of $2 * 10^{-4}$ and a batch size of 32. Training is conducted by using the auxiliary training objective until validation loss stabilizes.

For the language decoder, a Q-former is employed, comprising $K = 8$ query tokens, each with a dimension equivalent to LLaMA-7B's embedding dimension of 4096. $n = 8$ temporal tokens are selected, and noise injection is applied with a variance of $\sigma^2 = 10^{-4}$. In the first stage of fine-tuning, the Q-former undergoes fine-tuning using the Adam optimizer [42] with a learning rate of $2 * 10^{-6}$ on an NVIDIA T4 GPU for 20 epochs, during which the validation loss stabilizes. Moving to the second stage of fine-tuning, the LLaVA-Instruct dataset [51] is chosen, containing images from the Microsoft COCO [26] dataset alongside corresponding example conversations. ImageBind is utilized to generate visual embeddings for images in LLaVA-Instruct. Training in this stage follows the same settings as the first one, with the learning rate adjusted to $3 * 10^{-6}$ and conducted for 5000 iterations, leading to a qualitative assessment indicating the restoration of instructive capabilities.

## 5 EVALUATION

In this section, a comprehensive evaluation of Sensor2Text's capabilities is provided. We showcase sample conversations with Sensor2Text, demonstrating its proficiency in daily activity captioning, question answering, and reasoning. We also systematically compare Sensor2Text against similar models using quantitative metrics. Additionally, we explore Sensor2Text's proficiency in leveraging multiple modalities of sensor data, its robust ability to generalize to unseen users, and ablation studies to assess the contribution of various components of the model. Finally, we assess Sensor2Text's performance when applied on low information-density IMU sensors.

### 5.1 Dataset

We select the ActionSense [29] dataset, a multi-modal wearable sensor dataset for human activities in a kitchen environment, for use in the primary qualitative and quantitative evaluations of the model. The ActionSense dataset consists of 10 subjects, each performing kitchen activities in episodes spanning 60-120 minutes, such as preparing food, cleaning tableware, and interacting with items. In total, there are 20 diverse kitchen activities distributed between these categories. The full set of activities is depicted in Figure 5.

Various wearable sensor data modalities are collected during each episode and are paired with corresponding visual footage and text labels. From the available data modalities, we select the body-tracking, muscle activation, and eye-tracking due to them being consistently present for all episodes. The data is gathered using the Xsens MTw Awinda body tracking system [71], the Pupil Core eye-tracking headset from Pupil Labs [45], EMG sensors from Thalmic Lab's Myo Gesture Control Armband. Each wearable sensor records readings stored as sequential data of different frequencies. Additionally, we use visual footage collected from on-head cameras.

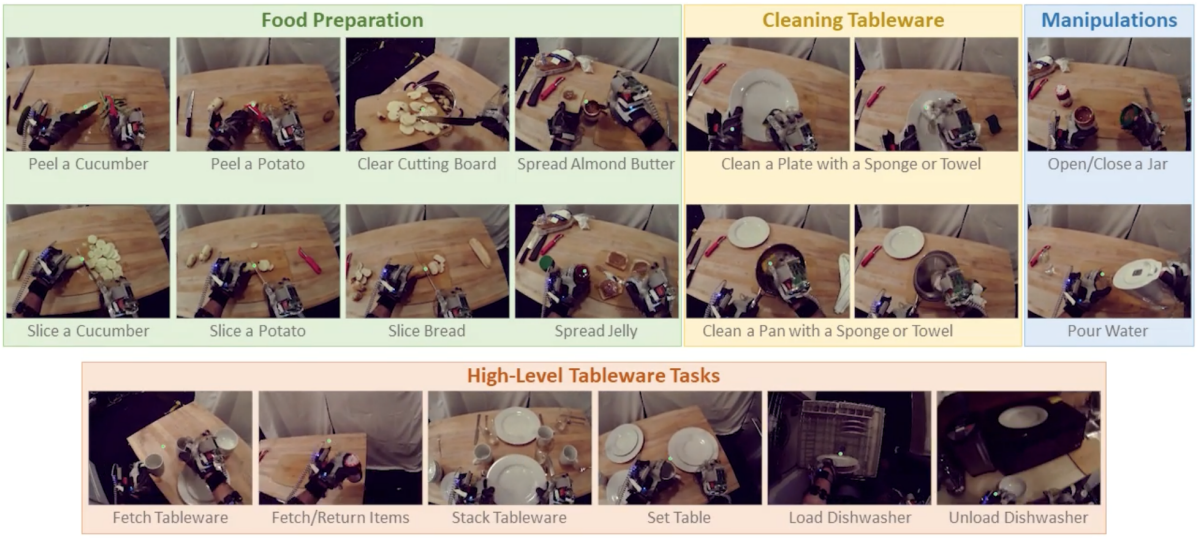



Fig. 5. ActionSense dataset. It consists of a variety of kitchen tasks, each with corresponding wearable sensor data, video data, and activity labels. Subjects perform a subset of these tasks over episodes of approximately 1 hour, with 10 total subjects.

*5.1.1 Data Preprocessing.* To ensure the strong performance and reliability of our model, we preprocess and normalize each modality of wearable sensor data. We sample each at a frequency of 50Hz. For eye-tracking gaze data, we remove outliers by clipping values outside the range of $[0.05, 0.95]$, handle missing or faulty values through interpolation, and normalize the data to the range $[-1, 1]$. For Electromyography (EMG) data, we apply a low-pass filter with a cutoff frequency of 5Hz to smooth the signal and remove high-frequency noise, followed by normalizing each data value to the range $[-1, 1]$. Lastly, for the body tracking data, we preprocess the joint rotation values by normalizing them to the range $[-1, 1]$, assuming that the joint rotation data falls within the range of $[-180, 180]$ degrees. These preprocessing steps handle potential faulty sensor readings, thus enhancing the input data's quality and improving our deep learning architecture's performance.

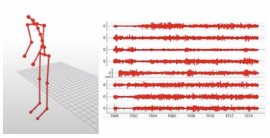
We partition the dataset into train, validation, and test subsets following a 70/15/15 split ratio, respectively. The episodes in ActionSense are segmented into 2-second clips to train the sensor encoder, which significantly increases the number of training examples available. For the language decoder, labeled sections of each episode are chosen, segmented into 16-second clips, and then trained using text generation loss. In total, there are 16000 2-second segments for the cross-modality supervised training of the sensor encoder and 1100 16-second segments that are adequately labeled. After incorporating a curated list of training prompts for the latter, there are approximately 7700 training examples for end-to-end text generation training.

## 5.2 Qualitative Evaluation

This section presents examples of Sensor2Text in diverse conversational contexts. For each instance, wearable sensor data from an unseen subject in the ActionSense dataset is utilized, and multiple questions are posed to the chatbot to assess its responses. For qualitative analysis, ground truth of the tested human activities is illustrated using footage from video cameras. These recordings are subsequently inputted into vision-language models to facilitate comparison with other models of the same category. It's worth noting that these visual data were not provided to Sensor2Text; the sensor-language model only receives the aforementioned wearable sensor data

**Sensor and Vision LLM Comparison**

| | |
|---|---|
| User | Describe what is happening in this scene. |
| VideoLLaMA | The person is cutting a cucumber. |
| User | Do they seem like they're good at cooking? |
| VideoLLaMA | They seem to be very skilled at slicing the cucumber. |
| User | Describe what is happening in this scene. |
| GPT4 | The image shows a person's forearms, wearing wrist guards with orange detailing, cutting a small, light-colored object on a wooden cutting board with a knife. The setting appears to be from a first-person perspective, possibly from a head-mounted camera. |
| User | Do they seem like they're good at cooking? |
| GPT4 | It's not clear from this one action whether they're good at cooking; we'd need to see more of their technique and the final dish to judge their cooking skills! |



| | |
|---|---|
| User | Describe what is happening in this scene. |
| Sensor2Text | A person is slicing a cucumber. |
| User | Do they seem like they're good at cooking? |
| Sensor2Text | The person seems to be very skilled at slicing the cucumber, as they are doing it with a lot of precision and care. |

Fig. 6. Comparison between Sensor2Text, GPT4, and fine-tuned VideoLLaMA. Each model demonstrates similar chat capabilities. However, Sensor2Text recalls the details of the subject's actions more precisely, demonstrating the advantages of using wearable sensors over video for some activities.

modalities as input. The following capabilities of Sensor2Text were evaluated and will be further discussed in the following subsections: sensor perception ability, conversational capabilities, and external knowledge integration. Additionally, a qualitative comparison between Sensor2Text and existing video language models is provided, including GPT4 [58], Claude Opus [3], and a fine-tuned open source model VideoLLaMA [77]. The performance of the sensor-language model is also compared against vision models using footage from inadequate lighting or unfavorable viewpoints, showing its superior performance in these settings.

*5.2.1 Sensor Perception Ability.* Figure 6 illustrates Sensor2Text's proficiency in identifying various activities from sensor data. Sensor2Text correctly identifies that the subject is slicing a cucumber and notes the subject's precise movements during the slicing process, showing its powerful sensor perception abilities.

Furthermore, in Figure 7, Sensor2Text accurately identifies that the subject is retrieving items from a cabinet and correctly identifies that the cabinet is above the subject. In contrast, all three video models are unable to determine any details of the activities due to the inadequate lighting. In Figure 8, during intervals with bad camera angles, the sensor model is again able to determine the subject's activity while vision models fail. This shows the superior performance of the sensor-based model in diverse settings.

*5.2.2 Conversational Capability.* Sensor2Text excels at interactive conversations with users, seamlessly blending the conversational abilities of LLaMA with insights derived from provided sensor data. Illustrated in Figure 6, when prompted, Sensor2Text correctly identifies that the subject is slicing a cucumber. When asked if the subject appears good at cooking, the model responds affirmatively and cites the subject's precise movements during the slicing process. This demonstrates Sensor2Text's ability to integrate the information from the inputted sensor data together with the language model's reasoning capabilities for smooth conversations. This integration enables dynamic and contextually aware interactions.

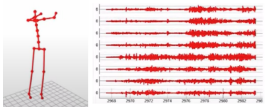
| **Sensor and Vision LLM Comparison, Poor Lighting** | | |
|---|---|---|
|  | User | Describe what is happening in this scene. |
| | VideoLLaMA | The scene is a man sitting on a couch, looking at a computer screen. (incorrect) |
| | GPT-4 | This image shows someone's knees bent at what appears to be a close-up angle, with the person wearing knee braces or supports. It seems like they are undergoing some sort of physical therapy or medical treatment, possibly using a machine designed for knee or leg exercises. The lighting is quite low, so the details are not very clear. (incorrect) |
| | Claude Opus | The image shows a person's legs and feet resting on what appears to be a desk or table. The room is dimly lit, with a green light visible near the top of the frame, possibly coming from a device like a modem or router. The person is wearing black shoes or sandals. Without being able to see more of the scene or identify the individual, there's not much additional context I can provide about what exactly is happening, but it seems to depict someone sitting at a desk or workstation in a dimly lit room or office. (incorrect) |
|  | User | Describe what is happening in this scene. |
| | Sensor2Text | A person is getting items from a cabinet. |
| | User | Is the cabinet above or below the person? |
| | Sensor2Text | The cabinet is above the person. (correct) |

Fig. 7. Comparison between Sensor2Text and vision-based models on dark examples. The sensor-based model correctly determines the subject's actions and the locations of objects in the scene. In contrast, the vision-based models are unable to determine the subject's activity.

*5.2.3 External Knowledge Integration.* Sensor2Text is capable of drawing upon world knowledge while integrating insights from provided sensor data. In Figure 8, Sensor2Text correctly assesses that it will take the subject about 30 seconds to peel the potato, showcasing its ability to utilize the information provided by the sensors and its inherent world knowledge. This enables dynamic and context-aware interactions.

## 5.3 Quantitative Evaluation

*5.3.1 Evaluation Metrics.* Providing a quantitative evaluation for Sensor2Text is challenging due to the unstructured nature of conversation-based tasks. While visual language models have established benchmarks like VQA [4], LLaVA-Bench [51], and ScienceQA [52], which assess conversational models' performance via meticulously crafted tasks, prompts, and metrics, comparable benchmarks are lacking for sensor language tasks. For example, Visual Question-Answering (VQA) assesses a model's correctness by comparing its output against a large set of human annotations [4]. In the absence of such benchmarks for sensor language tasks, we employ captioning scores as a proxy for evaluating performance, leveraging the availability of ground truth captions in our training data.

To assess Sensor2Text's performance, the model is prompted with a generic prompt requesting a caption. Such a prompt is illustrated in Figure 4. Then, traditional image captioning benchmarks are used to score the model's response. Following Microsoft COCO [26], we utilize the BLEU [59], ROUGE-L [50], METEOR [6], CIDEr [67], and SPICE [2] evaluation metrics. BLEU considers the co-occurrences of words between the predicted

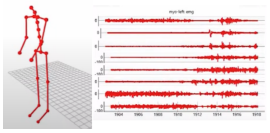
| **Sensor and Vision LLM Comparison, Unfavorable Viewpoints** | | |
|---|---|---|
|  | User | What is the person doing? |
| | VideoLLaMA | The person is cutting a cake with a knife. (incorrect) |
| | GPT-4 | The person in the image appears to be in the middle of a physical activity, possibly exercising or participating in some kind of sport. They are wearing what looks like athletic gear, including gloves and possibly wrist guards or fitness trackers on both arms. Due to the perspective of the image, which seems to be taken from a camera mounted on the person's head or chest, it's difficult to provide precise details, but it's clear that the person is engaged in a physical, probably sports-related, activity. (incorrect) |
| | Claude Opus | The image shows a person lying down, likely on a bed or couch. Their legs and feet are visible, wearing what appear to be black knee pads or braces. Without seeing more of the person or surrounding context, it's difficult to say exactly what they are doing, but they seem to be resting or relaxing based on their reclining position. (incorrect) |
|  | User | What is the person doing? |
| | Sensor2Text | A person is peeling a potato. |
| | User | How long do you think this will take? |
| | Sensor2Text | It will take about 30 seconds. (correct) |

Fig. 8. Comparison between Sensor2Text and vision-based models on an activity with a bad on-head camera angle. Again, Sensor2Text performs equally well and correctly determines the activity, while all of the vision-based models hallucinate details and are unable to answer correctly.

and ground truth sentences, with BLEU-n accounting for the co-occurrences of n words. ROUGE-L measures the longest common subsequence between the predicted and ground truth captions. METEOR computes the alignment between captions by considering exact token matches, stemming, and paraphrase matching, taking into account precision, recall, and word order. CIDEr assesses the similarity between a generated caption and a set of ground truth captions using TF-IDF weighting for each n-gram. SPICE evaluates the semantic similarity by extracting semantic propositions from the captions and computing the F1 score between the propositions of the predicted and ground truth captions. Together, these scores provide a comprehensive evaluation of the captioning performance of language models.

*5.3.2 Captioning.* The LLMs are prompted with a randomly selected prompt from a predefined list to elicit a caption output. This approach encourages the language model to produce outputs closely aligned with the terse, one-sentence descriptions of scenes found in the ActionSense dataset annotations.

Table 1 showcases the quantitative captioning results where the model undergoes evaluation against various state-of-the-art vision-based Language Models (LLMs). Specifically, a comparison is drawn with VideoLLaMA's Audio-Language (AL) branch [77], representing the omni-modal (video and audio) branch of VideoLLaMA and the source for the pre-trained Q-former fine-tuned for Sensor2Text. Additionally, comparisons are made against existing state-of-the-art closed-source models, GPT4 [58] and Claude Opus [3]. VideoLLaMA receives prompts based on the entire video, whereas GPT4 and Claude Opus are prompted with a representative frame sampled

from the video. Finally, we provide a comparison against a powerful visual captioning model, UniVL [53], which serves as a benchmark for video captioning.

To ensure a fair comparison between Sensor2Text and the selected vision-LLM models, additional techniques are employed to optimize their performance. For the open-source model VideoLLaMA, we apply additional fine-tuning by freezing the weights of the visual encoder and language decoder, and subsequently training the Q-former using text generation loss. This process follows a methodology similar to VideoLLaMA's original training procedure. The captioning scores for Sensor2Text and the fine-tuned VideoLLaMA are reported after stage 1 of finetuning, where the models are primed to generate the most direct and caption-like responses.

For the closed-source models, GPT4 and Claude Opus, a list of all activity labels is included in the prompt alongside the visual input and the LLM is repeatedly prompted until its response matches one of the provided labels. This partially compensates for the models' lack of finetuning on the dataset. Additionally, we provide captioning scores for VideoLLaMA using this prompting method without finetuning. All models are evaluated on the same test set as Sensor2Text, ensuring a consistent evaluation process.

Furthermore, benchmarking is conducted against existing video captioning models as well, specifically UniVL [53], after being fine-tuned on our dataset to facilitate an accurate comparison. It's important to note that multi-modal LLM models are disadvantaged when benchmarked against video-captioning models using captioning scores because captioning models are tuned to this specific task while multi-modal LLMs have powerful capabilities outside of captioning.

| Method | B@1 | B@2 | B@3 | B@4 | R | M | C | S |
|---|---|---|---|---|---|---|---|---|
| VideoLLaMA | 41.0 | 35.4 | 29.7 | 22.4 | 42.0 | 19.9 | 8.7 | 23.6 |
| Claude Opus | 64.3 | 57.7 | 51.6 | 44.9 | 65.7 | 33.7 | 33.9 | 48.8 |
| GPT4 | 72.3 | 67.8 | 63.8 | 59.4 | 75.7 | 42.3 | 54.0 | 64.6 |
| VideoLLaMA w/ finetuning | 80.6 | 76.5 | 73.5 | 70.4 | 81.4 | 48.7 | 66.2 | 73.0 |
| UniVL w/ finetuning | 82.6 | 78.5 | 74.8 | 70.9 | 83.1 | 49.0 | 62.5 | 67.5 |
| Sensor2Text | 78.7 | 74.5 | 71.1 | 67.5 | 80.5 | 46.7 | 62.2 | 68.6 |

Table 1. Captioning scores on the ActionSense dataset. We denote BLEU-n as B@n, ROUGE-L as R, METEOR as M, CIDEr as C, and SPICE as S. All scores are normalized for a full score of 100. Sensor2Text's performance is superior to un-finetuned vision LLM models and similar to that of finetuned vision models.

These scores illustrate how Sensor2Text achieves similar performance compared to both finetuned visual LLMs and finetuned video captioning models when generating captions on the ActionSense dataset and outperforms well-prompted closed-source vision LLMs. Remarkably, even with a relatively small number of training examples, our training methodology effectively leverages knowledge from visual-language models and extends it to sensor data. Additionally, as illustrated in Figure 7, the sensor-based model is unaffected by poor lighting conditions, in which vision-based models perform very poorly. These results validate the feasibility of using wearable sensors as an alternative to video for tracking daily activities.

*5.3.3 Generalization and Robustness.* This section showcases Sensor2Text's generalization capacity to unseen users. Due to the slight variations in data collected from each user, additional fine-tuning is often necessary for each new user. This process can impede the practical deployment of Sensor2Text, as collecting such data is time-consuming and arduous, presenting a challenge for new users.

A comparison is conducted between the model's performance when evaluated on data from users it has encountered before versus data from unseen users. Specifically, rather than splitting the train/validation/test sets by uniformly sampling across the entire dataset, partitioning by user is conducted instead. A subset of users is

selected for exclusion, and the model is trained on the remaining dataset for 20 epochs, mirroring the original model's training duration. Captioning performance is then evaluated on the held-out users. The captioning evaluation metrics comparing performance on data from seen users and unseen users are presented in Table 2.

| Method | B@1 | B@2 | B@3 | B@4 | R | M | C | S |
|---|---|---|---|---|---|---|---|---|
| unseen users | 73.8 | 69.6 | 66.2 | 62.5 | 77.7 | 43.3 | 58.9 | 68.0 |
| seen users | 78.7 | 74.5 | 71.1 | 67.5 | 80.5 | 46.7 | 62.2 | 68.6 |

Table 2. Captioning performance on the ActionSense dataset comparing totally unseen users versus seen users. Unseen users see a slight drop in performance, demonstrating Sensor2Text's powerful generalization capabilities.

Table 2 illustrates Sensor2Text's remarkable generalization capabilities and robustness. It exhibits only a minor decline in performance when applied to data from new users. This minimal degradation highlights the effectiveness of employing teacher-student networks and transfer learning in enhancing our model's generalization performance.

### 5.4 Multi-Modality

One crucial aspect of Sensor2Text lies in its usage of wearable sensor data across various modalities. This section examines the impact of multiple modalities on the model's training, contrasting it with the use of a single modality, accessed by conducting experiments where only a subset of data modalities is utilized. Specifically, several new models are trained using solely eye tracking, muscle activation, or body tracking data, following the same training procedure. The results are presented in the Table 3.

| Method | B@1 | B@2 | B@3 | B@4 | R | M | C | S |
|---|---|---|---|---|---|---|---|---|
| eye only | 52.2 | 44.4 | 38.2 | 31.0 | 52.4 | 24.3 | 17.4 | 34.1 |
| muscle only | 66.3 | 60.9 | 56.2 | 51.3 | 71.1 | 36.0 | 45.3 | 57.4 |
| body only | 73.4 | 67.9 | 63.6 | 59.1 | 74.5 | 42.2 | 53.4 | 62.9 |
| eye, muscle, body | 78.7 | 74.5 | 71.1 | 67.5 | 80.5 | 46.7 | 62.2 | 68.6 |

Table 3. Captioning scores achieved on ActionSense using only a subset of available sensor data modalities. Combining all modalities yields the best performance.

As demonstrated in the table, Sensor2Text shows stronger performance when given more modalities of sensor data, arising from the model's enhanced capability of differentiating between similar activities and mitigating the impact of noise in individual sensors. Such findings underscore Sensor2Text's effectiveness in leveraging diverse modalities of data, affirming our notion that the multiple data modalities of wearable sensor data provide advantages not present in video data.

### 5.5 Ablation Study

To show the necessity of the various components, experiments are conducted on the model's architecture and training process. Specifically, we demonstrate the need for the 2-step finetuning process, noise injection, and usage time-aware embeddings in the language decoder through ablation studies. The following alternatives are considered:

- Omitting language decoder fine-tuning
- Using a single embedding instead of 8 time-ordered embeddings
- Omitting noise injection

Each alternative is compared against the model using the same captioning scores as before. Consistency is maintained by keeping the sensor encoder and train/validation/test split constant. For the experiments that require re-training the language decoder, the same training procedures and applicable hyperparameters are employed as the baseline training. The results are shown in Table 4.

| Method | B@1 | B@2 | B@3 | B@4 | R | M | C | S |
|---|---|---|---|---|---|---|---|---|
| w/o finetuning | 41.8 | 35.2 | 28.7 | 19.0 | 44.3 | 17.1 | 7.2 | 22.5 |
| w/o temporal embeddings | 75.1 | 70.6 | 67.2 | 63.9 | 78.8 | 42.9 | 62.2 | 68.0 |
| w/o noise | 76.6 | 71.7 | 67.9 | 63.8 | 77.6 | 45.8 | 56.9 | 66.2 |
| Sensor2Text | 78.7 | 74.5 | 71.1 | 67.5 | 80.5 | 46.7 | 62.2 | 68.6 |

Table 4. Captioning results on the ActionSense dataset after omitting various components of the architecture.

As depicted in the table, Sensor2Text consistently exhibits the highest captioning scores across all evaluation metrics in all ablation studies, highlighting the improved performance achieved through finetuning, noise injection, and the integration of temporal tokens. In summary, these findings underscore the essential role of these components in optimizing the performance of the end-to-end model.

## 5.6 Low Information Density

Although Sensor2Text shows strong performance on the ActionSense dataset, the high information-density sensors present in the dataset may be difficult to collect in practice for daily activity tracking. For instance, collecting full body-tracking data involves wearing bulky sensors attached to a full-body suit. Thus, further investigation is needed to assess the applicability of Sensor2Text to sensors with low information density, such as smartphone and smartwatch IMUs. In this section, we select another dataset, the MMAct dataset [43], and provide a qualitative and quantitative evaluation to assess whether Sensor2Text can perform accurate Q&A and have meaningful conversations when applied to the MMAct dataset.

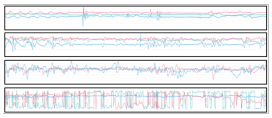
*5.6.1 MMAct Dataset.* To this end, we select the MMAct dataset [44]. In this dataset, 20 subjects perform activities from a set of 37 activities such as sitting, jumping, and entering/exiting a room. Wearable sensor data is collected from a smartphone and smartwatch, and the corresponding video is captured via environment-mounted video cameras. The smartphone collects acceleration, gyroscope, and orientation data, while the smartwatch further collects a set of smartwatch acceleration data for 4 total sensor data modalities. In contrast to the wearable sensor data collected from the ActionSense dataset, the sensors used in MMAct are more feasible to collect during everyday life but have lower information density.

To preprocess the data, we sample each modality of sensor data at a frequency of 50Hz. Each acceleration, gyroscope, and orientation signal is given in the three orthogonal directions $(x, y, z)$. We begin by computing the average magnitude across the dataset for each sensor modality and normalizing each to have an average magnitude of 1. The collected data may be sensitive to sensor placement; thus, following [44], we further compute $R_i = \arcsin\left(\frac{z_i}{\sqrt{x_i^2+y_i^2+z_i^2}}\right)$ for each data point and concatenate this to the original data. In total, each sensor data modality consists of 4 data values, significantly lower than the 66 data values from ActionSense's body tracking data or 16 from ActionSense's EMG data, signifying the lower information density. We use 60% of the data for training, 20% for validation, and 20% for testing. The same training settings are applied as before, with for 200 epochs for the sensor encode, 20 epochs for the language decoder trained, and 5000 iterations for instruction finetuning.

*5.6.2 Qualitative Evaluation.* This section presents examples of Sensor2Text applied to activities from the MMAct dataset. Again, a comparison is presented with existing vision language models, including state-of-the-art closed-source LLMs. Sensor2Text is provided all 4 modalities of wearable sensor data but not the visual data. To facilitate an efficient comparison, we further configure a system prompt for the vision-language model to help the model filter out details such as the labels on the walls and instead focus on the subject. This ensures that its response is concise and focuses on details useful to our evaluation of daily activity perception ability. We compare and discuss Sensor2Text's performance against the qualitative results from ActionSense.

**Sensor and Vision LLM Comparison, IMU sensors, high-level scene understanding**

| | |
|---|---|
| User | What is happening in this scene? |
| GPT-4 | The subject is standing near a wall, possibly adjusting or interacting with a device or panel on the wall. They appear to be stationary and focused on the task involving the wall-mounted object. (incorrect) |
| User | Does the person seem relaxed or in a hurry? |
| GPT-4 | The person appears relaxed, as there are no visible signs of haste or urgency in their posture or movements. (correct) |



| | |
|---|---|
| User | What is happening in this scene? |
| Sensor2Text | A person is entering the room. (correct) |
| User | Does the person seem relaxed or in a hurry? |
| Sensor2Text | The person seems to be in a hurry. (incorrect) |

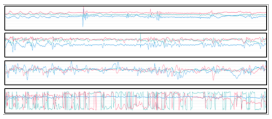
Fig. 9. Comparison between Sensor2Text and GPT-4 on the MMAct dataset, comparing both models' understanding of the scene. The subject cautiously enters the room from outside. Sensor2Text is able to determine the subject's activity from the movement data gathered by IMU sensors. However, it incorrectly answers that the person is in a hurry. In contrast, GPT-4 is unable to determine the subject's action because of its lack of temporal understanding but correctly answers that the person is relaxed.

As shown in Figures 9 and 10, Sensor2Text is able to leverage the smartphone/smartwatch sensors to correctly determine the activity being performed. In Figure 9, Sensor2Text correctly determines that the subject is entering the room, demonstrating its ability to function in low information density settings. However, Sensor2Text often fails to understand further details. In the same figure, when asked whether the subject appears to be in a hurry, Sensor2Text incorrectly replies yes, even though the subject is being notably cautious and moving slowly. In contrast, GPT-4 is unable to determine the subject's actions due to the lack of temporal context in images but is able to recognize that the subject is relaxed. In comparison to GPT-4, Sensor2Text applied on IMU sensors is more capable of daily activity tracking but less capable of grasping details. In Figure 10, Sensor2Text is able to correctly determine that the subject is picking an item off the ground. However, when asked to describe details of the item, Sensor2Text is only able to correctly reply that the item is light but hallucinates other details about the item. This demonstrates that Sensor2Text is only able to use smartphone and smartwatch sensors to deduce a limited set of information about the activity due to the inherent limitations of the sensors themselves.

*5.6.3 Quantitative Evaluation.* A quantitative evaluation is further conducted on Sensor2Text, with captioning scores serving as a proxy to compare Sensor2Text's performance against existing vision language models. Following the same procedure as before, the list of possible captions to the prompt is prepended when prompting

**Sensor and Vision LLM Comparison, IMU sensors, detail recall**

| | |
|---|---|
| User | What is the person doing? |
| GPT-4 | The subject in the scene is crouching down and appears to be picking something up or placing something on the floor. |
| User | Describe the item they are picking up. What is its shape and size? Does it seem heavy or light? |
| GPT-4 | The item the subject is picking up is small and likely flat, given the way it is being handled. It seems to be light, as the subject is using only one hand to pick it up with ease. (correct) |



| | |
|---|---|
| User | What is the person doing? |
| Sensor2Text | A person is picking something up. (correct) |
| User | Describe the item they are picking up. What is its shape and size? Does it seem heavy or light? |
| Sensor2Text | The item is a phone. It is light in weight and has a round shape. (incorrect) |

Fig. 10. Comparison of Sensor2Text and GPT-4 on the MMAct dataset, focusing on both models' ability to recall details. Both models are able to determine the subject's actions. GPT-4 is further able to estimate the properties of the object on the floor, whereas Sensor2Text hallucinates.

closed-source vision LLMs to ensure that their outputs are in the target distribution. The results are given in Table 5. Again, Sensor2Text achieves higher captioning scores than vision LLM models without finetuning. This demonstrates Sensor2Text's remarkable ability to leverage low information-density sensors to determine the subject's activity. However, as demonstrated in the qualitative evaluation, Sensor2Text's performance in detail recognition decreases when applied to low information-density sensors. This shows that a decrease in the information density primarily causes a decrease in detail recall performance, whereas the model can still successfully leverage low information-density sensors to achieve activity recognition.

| Method | B@1 | B@2 | B@3 | B@4 | R | M | C | S |
|---|---|---|---|---|---|---|---|---|
| VideoLLaMA | 52.5 | 46.2 | 38.3 | 19.5 | 56.2 | 30.2 | 2.8 | 31.9 |
| Claude Opus | 62.9 | 58.1 | 51.2 | 35.7 | 64.9 | 30.4 | 11.2 | 49.6 |
| GPT4 | 70.5 | 67.2 | 62.9 | 54.2 | 75.3 | 36.6 | 29.6 | 61.7 |
| VideoLLaMA w/ finetuning | 78.0 | 74.0 | 69.4 | 61.8 | 79.2 | 42.9 | 45.1 | 70.4 |
| Sensor2Text | 73.6 | 68.8 | 63.1 | 53.3 | 74.7 | 38.3 | 33.3 | 62.4 |

Table 5. Captioning scores on the MMAct dataset. Sensor2Text's performance is superior to that of un-finetuned vision LLM models and comparable to that of finetuned models, demonstrating its remarkable ability to use low-information density sensors such as IMU.

## 5.7 Privacy

One advantage of using sensors to perform daily activity tracking is the preservation of users' privacy. Figure 11 presents a visualization of the decrease in information density between video and various sensor data modalities.

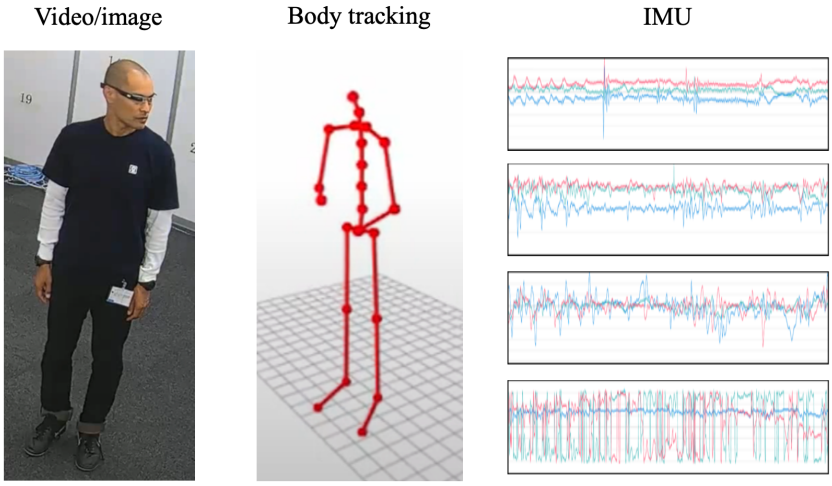



Fig. 11. Visualization of information associated with each data modality. Vision captures the most, including subjects' identifiable facial information, clothing, and surroundings. Body tracking captures less information, while IMU sensors capture the least.

As shown, while video captures identifiable information about a subject, wearable sensors are much more privacy-preserving. For example, in one survey [27] regarding the privacy of wearable sensors for daily activity tracking and healthcare monitoring, video-based monitoring is identified as more invasive of a subject's surroundings, whereas many privacy concerns regarding wearable sensors can be mitigated with proper data handling and anonymization.

The evaluation using eye-tracking only in ActionSense or privacy-preserving sensors from the MMAct dataset demonstrates that while the model can leverage low-information density sensors such as smartphone and smartwatch IMU to identify users' basic activities, it has reduced performance on detail recognition and recall. Using low information-density sensors would suffice for everyday use, such as inquiring about when a subject sleeps, exercises, or is working on a keyboard. In contrast, for healthcare applications where precision is required, such as inquiring about the type of pill bottle a subject has picked up, high information-density sensors are ideal. Thus, the trade-off between privacy and performance should be carefully considered when deploying such a model for daily activity tracking.

## 6 LIMITATIONS AND FUTURE WORK

Our model demonstrates strong performance when applied to high information-density sensors such as full body-tracking and EMG data than IMU data. However, these can be cumbersome to collect for practical daily use. Additionally, compared to smartphones and smartwatches IMU, full body-tracking data is more invasive of privacy. Thus, the trade-off in performance between employing sensors containing higher information density versus privacy-preserving sensors should be carefully considered when deploying such a model to daily activity tracking. Future work could investigate supplementing smartphone and smartwatch sensors with other privacy-preserving sensors to compensate for the reduced information density while still ensuring strong daily activity-tracking capabilities. Still, we contend that even full-body tracking and EMG sensors are more privacy-preserving than constant video surveillance, as demonstrated in Figure 11.

Secondly, while the use of foundational vision models and pre-trained large language models greatly enhances the training effectiveness and performance of the Sensor2Text, it also limits the applicability of the model. The use of vision models during training restricts the applicable datasets to those that include vision data for training, which limits the applicability of the model. Furthermore, the use of large language models makes Sensor2Text more computationally expensive compared to classification or captioning models and presents a bottleneck for inference speed. Finally, the applications of the model depend on the licenses of the employed components. In particular, LLaMA 2's community license specifies limits on its commercial use for products with large monthly user counts, and ImageBind forbids commercial use altogether. These can be remedied by selecting an alternative vision model or large language model but nonetheless can present difficulties for the adoption of Sensor2Text. Thus, further work could investigate novel transfer learning methods that potentially reduce the reliance of a sensor-language model on existing vision-language models.

Finally, while Sensor2Text shows the potential to generalize to new users and can be applied in various environments, it struggles with completely unseen environments or activity types. For example, vision-language models like VideoLLaMA and GPT-4 can handle visual footage from new environments without additional training, but Sensor2Text requires further training due to challenges in sensor data collection, such as sensor placements and device variations. The relatively small size of existing sensor-based datasets for daily activity tracking additionally increases the difficulty for a sensor-based model to generalize to new activities. Nonetheless, the strong performance of Sensor2Text with minimal training data suggests it could be a promising alternative to traditional video-based daily activity tracking, addressing privacy and viewpoint concerns. Future research could explore the collection of a large-scale sensor and video dataset for daily activity tracking to train a sensor-language model that can generalize to entirely new activity types.

## 7 CONCLUSION

In conclusion, this paper introduces a novel, sensor-based approach to enable question-answering on tracking a person's daily activities. Our encoder-decoder model adeptly integrates multiple sensor modalities and text to generate natural language responses, achieving a capability previously only attainable with visual data or a limited number of other modalities to a constrained degree. Despite certain challenges, particularly the scarcity of sensor data and paired sensor/text data, extensive evaluation demonstrates our model's excellent performance across various natural language tasks conditioned on sensor data. This pioneering work highlights the potential of using wearable sensors as an alternative to video in applications such as tracking daily activities, opening up new possibilities for privacy-preserving and unobtrusive monitoring systems.

## REFERENCES

[1] Fadel Adib, Zachary Kabelac, Dina Katabi, and Robert C. Miller. 2014. 3D tracking via body radio reflections. In *Proceedings of the 11th USENIX Conference on Networked Systems Design and Implementation* (Seattle, WA) *(NSDI'14)*. USENIX Association, USA, 317–329.

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. arXiv:1607.08822 [cs.CV]

[3] Anthropic. 2024. Claude 3 Family. https://www.anthropic.com/news/claude-3-family. Accessed: 2024-04-19.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

[5] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer Normalization. arXiv:1607.06450 [stat.ML]

[6] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss (Eds.). Association for Computational Linguistics, Ann Arbor, Michigan, 65–72. https://aclanthology.org/W05-0909

[7] Abdelkareem Bedri, Richard Li, Malcolm Haynes, Raj Prateek Kosaraju, Ishaan Grover, Temiloluwa Prioleau, Min Yan Beh, Mayank Goel, Thad Starner, and Gregory Abowd. 2017. EarBit: Using Wearable Sensors to Detect Eating Episodes in Unconstrained Environments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 37 (sep 2017), 20 pages. https://doi.org/10.1145/3130902

[8] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL]

[9] Youngjae Chang, Akhil Mathur, Anton Isopoussu, Junehwa Song, and Fahim Kawsar. 2020. A Systematic Study of Unsupervised Domain Adaptation for Robust Human-Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1, Article 39 (mar 2020), 30 pages. https://doi.org/10.1145/3380985

[10] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. 2023. VAST: A Vision-Audio-Subtitle-Text Omni-Modality Foundation Model and Dataset. arXiv:2305.18500 [cs.CV]

[11] Wenqiang Chen, Daniel Bevan, and John Stankovic. 2021. ViObject: A Smartwatch-based Object Recognition System via Vibrations. In *Adjunct Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology*. 97–99.

[12] Wenqiang Chen, Lin Chen, Yandao Huang, Xinyu Zhang, Lu Wang, Rukhsana Ruby, and Kaishun Wu. 2019. Taprint: Secure text input for commodity smart wristbands. In *The 25th Annual International Conference on Mobile Computing and Networking*. 1–16.

[13] Wenqiang Chen, Lin Chen, Meiyi Ma, Farshid Salemi Parizi, Shwetak Patel, and John Stankovic. 2021. ViFin: Harness Passive Vibration to Continuous Micro Finger Writing with a Commodity Smartwatch. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–25.

[14] Wenqiang Chen, Lin Chen, Meiyi Ma, Farshid Salemi Parizi, Patel Shwetak, and John Stankovic. 2020. Continuous micro finger writing recognition with a commodity smartwatch: demo abstract. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 603–604.

[15] Wenqiang Chen, Lin Chen, Kenneth Wan, and John Stankovic. 2020. A smartwatch product provides on-body tapping gestures recognition: demo abstract. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 589–590.

[16] Wenqiang Chen, Maoning Guan, Yandao Huang, Lu Wang, Rukhsana Ruby, Wen Hu, and Kaishun Wu. 2018. Vitype: A cost efficient on-body typing system through vibration. In *2018 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 1–9.

[17] Wenqiang Chen, Maoning Guan, Yandao Huang, Lu Wang, Rukhsana Ruby, Wen Hu, and Kaishun Wu. 2019. A Low Latency On-Body Typing System through Single Vibration Sensor. *IEEE Transactions on Mobile Computing* 19, 11 (2019), 2520–2532.

[18] Wenqiang Chen, Maoning Guan, Lu Wang, Rukhsana Ruby, and Kaishun Wu. 2017. FLoc: Device-free passive indoor localization in complex environments. In *2017 IEEE International Conference on Communications (ICC)*. IEEE, 1–6.

[19] Wenqiang Chen, Yexin Hu, Wei Song, Yingcheng Liu, Antonio Torralba, and Wojciech Matusik. 2024. CAvatar: Real-time Human Activity Mesh Reconstruction via Tactile Carpets. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 4 (2024), 1–24.

[20] Wenqiang Chen, Yanming Lian, Lu Wang, Rukhsana Ruby, Wen Hu, and Kaishun Wu. 2017. Virtual keyboard for wearable wristbands. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*. 1–2.

[21] Wenqiang Chen, Shupei Lin, Zhencan Peng, Farshid Salemi Parizi, Seongkook Heo, Shwetak Patel, Wojciech Matusik, Wei Zhao, and John Stankovic. 2024. ViObject: Harness Passive Vibrations for Daily Object Recognition with Commodity Smartwatches. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 1, Article 5 (mar 2024), 26 pages. https://doi.org/10.1145/3643547

[22] WENQIANG CHEN, SHUPEI LIN, ELIZABETH THOMPSON, and JOHN STANKOVIC. 2021. SenseCollect: We Need Efficient Ways to Collect On-body Sensor-based Human Activity Data! *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol* 1, 1 (2021).

[23] Wenqiang Chen and John Stankovic. 2022. ViWatch: harness vibrations for finger interactions with commodity smartwatches. In *Proceedings of the 13th ACM Wireless of the Students, by the Students, and for the Students Workshop*. 4–6.

[24] Wenqiang Chen, Ziqi Wang, Pengrui Quan, Zhencan Peng, Shupei Lin, Mani Srivastava, Wojciech Matusik, and John Stankovic. 2023. Robust Finger Interactions with COTS Smartwatches via Unsupervised Siamese Adaptation. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–14.

[25] Wenqiang Chen, Ziqi Wang, Pengrui Quan, Zhencan Peng, Shupei Lin, Mani Srivastava, and John Stankovic. 2022. Making Vibration-based On-body Interaction Robust. In *2022 ACM/IEEE 13th International Conference on Cyber-Physical Systems (ICCPS)*. IEEE, 300–301.

[26] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. arXiv:1504.00325 [cs.CV]

[27] Prerit Datta, Akbar Siami Namin, and Moitrayee Chatterjee. 2018. A Survey of Privacy Concerns in Wearable Devices. In *2018 IEEE International Conference on Big Data (Big Data)*. 4549–4553. https://doi.org/10.1109/BigData.2018.8622110

[28] Shohreh Deldari, Hao Xue, Aaqib Saeed, Daniel V. Smith, and Flora D. Salim. 2022. COCOA: Cross Modality Contrastive Learning for Sensor Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 108 (sep 2022), 28 pages. https://doi.org/10.1145/3550316

[29] Joseph DelPreto, Chao Liu, Yiyue Luo, Michael Foshey, Yunzhu Li, Antonio Torralba, Wojciech Matusik, and Daniela Rus. 2022. ActionSense: A Multimodal Dataset and Recording Framework for Human Activities Using Wearable Sensors in a Kitchen Environment. In *Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*. https://action-sense.csail.mit.edu

[30] Lijie Fan, Tianhong Li, Yuan Yuan, and Dina Katabi. 2020. In-Home Daily-Life Captioning Using Radio Signals. arXiv:2008.10966 [cs.CV]
[31] Nadine Fligge, Holger Urbanek, and Patrick van der Smagt. 2013. Relation between object properties and EMG during reaching to grasp. *Journal of Electromyography and Kinesiology* 23, 2 (2013), 402–410. https://doi.org/10.1016/j.jelekin.2012.10.010
[32] Zohreh Ghaderi, Leonard Salewski, and Hendrik P. A. Lensch. 2022. Diverse Video Captioning by Adaptive Spatio-temporal Attention. arXiv:2208.09266 [cs.CV]
[33] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. ImageBind: One Embedding Space To Bind Them All. arXiv:2305.05665 [cs.CV]
[34] Maoning Guan, Wenqiang Chen, Yandao Huang, Rukhsana Ruby, and Kaishun Wu. 2019. FaceInput: a hand-free and secure text entry system through facial vibration. In *2019 16th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 1–9.
[35] Yu Guan and Thomas Plötz. 2017. Ensembles of Deep LSTM Learners for Activity Recognition using Wearables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 2, Article 11 (jun 2017), 28 pages. https://doi.org/10.1145/3090076
[36] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, and Pheng-Ann Heng. 2023. Point-Bind & Point-LLM: Aligning Point Cloud with Multi-modality for 3D Understanding, Generation, and Instruction Following. arXiv:2309.00615 [cs.CV]
[37] Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, Xudong Lu, Shuai Ren, Yafei Wen, Xiaoxin Chen, Xiangyu Yue, Hongsheng Li, and Yu Qiao. 2023. ImageBind-LLM: Multi-modality Instruction Tuning. arXiv:2309.03905 [cs.MM]
[38] Harish Haresamudram, Irfan Essa, and Thomas Plötz. 2021. Contrastive Predictive Coding for Human Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 2, Article 65 (jun 2021), 26 pages. https://doi.org/10.1145/3463506
[39] Yandao Huang, Wenqiang Chen, Hongjie Chen, Lu Wang, and Kaishun Wu. 2019. G-fall: device-free and training-free fall detection with geophones. In *2019 16th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 1–9.
[40] 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:2401.06866 [cs.CL] https://arxiv.org/abs/2401.06866
[41] Evan King, Haoxiang Yu, Sangsu Lee, and Christine Julien. 2024. Sasha: Creative Goal-Oriented Reasoning in Smart Homes with Large Language Models. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 1, Article 12 (mar 2024), 38 pages. https://doi.org/10.1145/3643505
[42] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG]
[43] Quan Kong, Ziming Wu, Ziwei Deng, Martin Klinkigt, Bin Tong, and Tomokazu Murakami. 2019. MMAct: A large-scale dataset for cross modal human action understanding. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 8657–8666. https://doi.org/10.1109/ICCV.2019.00875
[44] Quan Kong, Ziming Wu, Ziwei Deng, Martin Klinkigt, Bin Tong, and Tomokazu Murakami. 2019. MMAct: A Large-Scale Dataset for Cross Modal Human Action Understanding. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 8657–8666. https://doi.org/10.1109/ICCV.2019.00875
[45] Pupil Labs. 2022. Pupil Core eye-tracking headset. https://pupil-labs.com/products/core.
[46] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597 [cs.CV]
[47] Tianhong Li, Lijie Fan, Mingmin Zhao, Yingcheng Liu, and Dina Katabi. 2019. Making the Invisible Visible: Action Recognition Through Walls and Occlusions. arXiv:1909.09300 [cs.CV]
[48] Xiang Li, Daqing Zhang, Qin Lv, Jie Xiong, Shengjie Li, Yue Zhang, and Hong Mei. 2017. IndoTrack: Device-Free Indoor Human Tracking with Commodity Wi-Fi. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 72 (sep 2017), 22 pages. https://doi.org/10.1145/3130940
[49] Yuanzhi Liang, Linchao Zhu, Xiaohan Wang, and Yi Yang. 2024. IcoCap: Improving Video Captioning by Compounding Images. *IEEE Transactions on Multimedia* 26 (2024), 4389–4400. https://doi.org/10.1109/TMM.2023.3322329
[50] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013
[51] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. arXiv:2304.08485 [cs.CV]
[52] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.
[53] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. 2020. UniVL: A Unified Video and Language Pre-Training Model for Multimodal Understanding and Generation. arXiv:2002.06353 [cs.CV]
[54] Haojie Ma, Zhijie Zhang, Wenzhong Li, and Sanglu Lu. 2021. Unsupervised Human Activity Representation Learning with Multi-task Deep Clustering. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 48 (mar 2021), 25 pages. https://doi.org/10.1145/3448074

[55] Shenghuan Miao, Ling Chen, Rong Hu, and Yingsong Luo. 2022. Towards a Dynamic Inter-Sensor Correlations Learning Framework for Multi-Sensor-Based Wearable Human Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 130 (sep 2022), 25 pages. https://doi.org/10.1145/3550331

[56] Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. 2021. Large Dual Encoders Are Generalizable Retrievers. arXiv:2112.07899 [cs.IR]

[57] Runliang Niu, Jindong Li, Shiqi Wang, Yali Fu, Xiyu Hu, Xueyuan Leng, He Kong, Yi Chang, and Qi Wang. 2024. ScreenAgent: A Vision Language Model-driven Computer Control Agent. arXiv:2402.07945 [cs.HC]

[58] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[59] Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. (10 2002). https://doi.org/10.3115/1073083.1073135

[60] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV]

[61] Jorge Reyes-Ortiz, Davide Anguita, Alessandro Ghio, Luca Oneto, and Xavier Parra. 2012. Human Activity Recognition Using Smartphones. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C54S4K.

[62] Jie Su, Zhenyu Wen, Tao Lin, and Yu Guan. 2022. Learning Disentangled Behaviour Patterns for Wearable-based Human Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 1, Article 28 (mar 2022), 19 pages. https://doi.org/10.1145/3517252

[63] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. arXiv:1908.07490 [cs.CL]

[64] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL]

[65] M.Z. Uddin and A. Soylu. 2021. Human activity recognition using wearable sensors, discriminant analysis, and long short-term memory-based neural structured learning. *Sci Rep* 11 (2021), 16455. https://doi.org/10.1038/s41598-021-95947-y

[66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. arXiv:1706.03762 [cs.CL]

[67] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based Image Description Evaluation. arXiv:1411.5726 [cs.CV]

[68] Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to Sequence – Video to Text. arXiv:1505.00487 [cs.CV]

[69] Yilin Wang, Suhang Wang, Jiliang Tang, Neil O'Hare, Yi Chang, and Baoxin Li. 2016. Hierarchical Attention Network for Action Recognition in Videos. arXiv:1607.06416 [cs.CV]

[70] Kaishun Wu, Yandao Huang, Wenqiang Chen, Lin Chen, Xinyu Zhang, Lu Wang, and Rukhsana Ruby. 2020. Power saving and secure text input for commodity smart watches. *IEEE Transactions on Mobile Computing* 20, 6 (2020), 2281–2296.

[71] Xsens. 2022. MTw Awinda wearable body-tracking system. https://www.xsens.com/products/mtw-awinda.

[72] Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, and Fei Huang. 2021. E2E-VLP: End-to-End Vision-Language Pre-training Enhanced by Visual Learning. arXiv:2106.01804 [cs.CV]

[73] Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, Guohai Xu, Ji Zhang, Songfang Huang, Fei Huang, and Jingren Zhou. 2023. mPLUG-2: A Modularized Multi-modal Foundation Model Across Text, Image and Video. arXiv:2302.00402 [cs.CV]

[74] Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2024. Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 1, Article 31 (mar 2024), 32 pages. https://doi.org/10.1145/3643540

[75] Hanhua Ye, Guorong Li, Yuankai Qi, Shuhui Wang, Qingming Huang, and Ming-Hsuan Yang. 2022. Hierarchical Modular Network for Video Captioning. arXiv:2111.12476 [cs.CV]

[76] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A Survey on Multimodal Large Language Models. arXiv:2306.13549 [cs.CV]

[77] Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. arXiv:2306.02858 [cs.CL]

[78] Mingmin Zhao, Yonglong Tian, Hang Zhao, Mohammad Abu Alsheikh, Tianhong Li, Rumen Hristov, Zachary Kabelac, Dina Katabi, and Antonio Torralba. 2018. RF-based 3D skeletons. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication* (Budapest, Hungary) *(SIGCOMM '18)*. Association for Computing Machinery, New York, NY, USA, 267–281. https://doi.org/10.1145/3230543.3230579