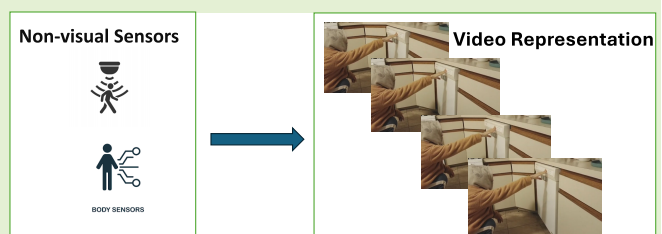


GeXSe (Generative Explanatory Sensor System): A Deep Generative Method for Human Activity Recognition of Smart Spaces IOT

Sun Yuan^{ID}, Member, IEEE, Salami Pargoo Navid^{ID}, and Ortiz Jorge^{ID}

Abstract—Effective sensemaking is crucial across various domains, where trust in system outputs is essential, highlighting a significant challenge in relying solely on sensor-based activity recognition. This limitation underscores the need for interpretable models that process raw data and provide comprehensible insights. Our work introduces generative explanatory sensor system (GeXSe), a novel multitask framework that jointly models raw sensors for classification while also generating grounded visual explanations. At the core of GeXSe lies a parallel multibranch multilayer perceptron fast Fourier convolution (PMB-MLP-FFC) module tailored for multimodal sensor fusion and explanation generation. PMB-MLP-FFC extracts interpretable features optimized for both tasks through multibranch parallel convolutions and Fourier transforms. We validated GeXSe across three diverse public datasets of daily activities recorded by camera, microphone, motion, and environmental sensors. Results show superior activity recognition over baseline models. Furthermore, human evaluation studies confirm the generated visual explanations enhance understanding and trust compared to purely sensor-based outputs.

Index Terms—Activity recognition, human activity recognition (HAR), sensor systems.



I. INTRODUCTION

THE rapid advancement of deep learning techniques has revolutionized human activity recognition (HAR), enabling highly accurate and complex models for interpreting human behaviors from sensor data. As these models become more sophisticated, there is a growing need for interpretability and explainability to ensure trust and transparency in HAR

Received 1 January 2025; revised 29 January 2025; accepted 7 February 2025. Date of publication 19 February 2025; date of current version 2 April 2025. This work was supported in part by the National Science Foundation (NSF) under the Center for Smart Streetscapes through NSF Cooperative under Agreement EEC-2133516; and in part by the DataCity Smart Mobility Testing Ground through Middlesex County Resolution 21-821-R, New Jersey Department of Transportation and Federal Highway Administration Research, under Project 21-60168. The associate editor coordinating the review of this article and approving it for publication was Dr. David Plets. (Corresponding author: Sun Yuan.)

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Rutgers University Office for Research under Application No. Pro2022000705.

The authors are with the School of Engineering, ECE Department, Winlab, Rutgers, The State University of New Jersey, New Brunswick, NJ 08901 USA (e-mail: ys820@soe.rutgers.edu; navid.salampargoo@rutgers.edu; jorge.ortiz@rutgers.edu).

Digital Object Identifier 10.1109/JSEN.2025.3541158

systems. Without understanding how these models make decisions and provide meaningful explanations, their adoption in critical applications, such as healthcare and smart environments, may be hindered.

Interpretability involves comprehending the internal mechanisms of a model and/or gaining insights into the conclusions drawn by that model. It focuses on the transparency of the model's structure and decision-making process or result. On the other hand, explainability is concerned with understanding why a model made a specific prediction for a given input, attributing the output back to the relevant input features [81]. While interpretability deals with the overall logic of the model, explainability provides insights into individual predictions [50], [76]. Both aspects are crucial in HAR applications, where trust in the system's outputs is paramount, and users need to comprehend the reasoning behind the model's decisions.

Example of interpretability: Consider a HAR system using a decision tree to classify human activities based on sensor data. Each node in the tree represents a decision based on input from accelerometers, gyroscopes, and environmental sensors. For instance, a node might split activities based on accelerometer

variance, with lower variances indicating stationary activities and higher variances suggesting movement. Further nodes differentiate between specific activities using gyroscope and environmental sensor patterns and thresholds.

Interpretability is achieved through the transparency of the decision tree structure. Users can comprehend the model's internal mechanisms by following the decision paths for classifying an activity. For example, if the model classifies an activity as "running," users can trace the decision tree to understand that this classification was based on high accelerometer variance, a specific gyroscope speed threshold, and environmental factors aligning with typical running conditions.

Example of explainability: Explainability is demonstrated when the HAR system provides a detailed explanation for a specific prediction. For instance, when the system classifies a segment of sensor data as "jogging," it might provide the following explanation: "This prediction was made because the accelerometer showed consistent, moderate variance indicative of repetitive motion, the gyroscope detected a steady forward movement at a jogging pace, and the environmental sensors detected an outdoor setting with moderate temperature and GPS data indicating a path through a park, which aligns with typical jogging behavior."

This explanation attributes the prediction to relevant input features from the sensor data, providing a clear rationale that links the data characteristics to the model's conclusion. Explainability increases trust in the system's decision-making process by offering insights into why the model made a particular prediction. It allows users to verify the reasoning against their domain knowledge.

Traditional HAR approaches, such as rule-based systems and classical machine learning algorithms like decision trees and linear regression, offer inherent interpretability due to their simple and transparent structures. However, these methods often struggle to capture the complex patterns and relationships in high-dimensional sensor data [58], which can be difficult to interpret due to the combination of multiple sensors, such as accelerometers, gyroscopes, and environmental sensors, each generating unique patterns and relationships [86]. Distinguishing between ambiguous activities, such as "sitting" and "standing," or "jogging" and "running," requires the model to capture subtle differences in sensor data. While the field has witnessed a paradigm shift with the adoption of deep learning techniques, offering robust solutions to this challenge [12], [13], [26], [51], [84], the black-box nature of these models makes them difficult to interpret and explain [4], hindering their adoption in critical applications where accountability and user trust are essential [6], [14], [21], [59].

Recent studies [85], [94] introduced saliency maps and feature importance scores to address the interpretability challenge in deep-learning-based HAR. However, these approaches often generate abstract visualizations or numerical values that may not be intuitive or actionable for users without machine learning expertise [3]. This lack of comprehensible explanations can lead to a disconnect between the model's outputs and the users' understanding, ultimately impacting the trust and acceptance of HAR systems. Furthermore, most research on sensor data

interpretation has focused on generating text-based, semantic explanations [5], [7], [34], [37], [61], [67], [91], with a notable gap in the use of vision-based, pictorial representations [33].

Deep generative models have emerged as a promising solution for deriving interpretable insights from complex sensor data. They bridge the gap between the technical complexity of sensor-based HAR and the provision of clear, actionable insights that resonate with human experiences.

A. Knowledge Gap

The growing interest in interpretable and explainable HAR models that rely on vision domain features is propelled by its potential to bolster the accuracy of activity recognition models [28]. Despite various approaches [5], [23], [34], [45], [79], the majority of current techniques for sensor-based HAR interpretations and explanations hinge on abstract visualizations, text descriptions, or labels, which may not always be intuitive or actionable for end-users. There is a need for explanation techniques that can effectively translate complex sensor data into *visually grounded* and *easily comprehensible* explanations, catering to end-users' needs in various domains [5], [33], [34].

To address the challenge of translating complex sensor data into meaningful and understandable insights for end-users, a comprehensive and user-centric explainability framework is essential. This framework should focus on bridging the gap between the intricate patterns in sensor data and the user's domain knowledge, ensuring that the explanations provided by the HAR system are not only accurate but also intuitive, actionable, and aligned with the user's understanding of the activities and context. Such a framework would not only facilitate the extraction of features from sensor data but also align these features with vision domain descriptions, thereby enabling a deeper understanding of model outcomes [5], [23], [34], [45], [79]. This alignment is vital for creating explicative models that can offer explanations intelligible to humans, fostering a genuine sense of explainability.

Moreover, a critical challenge is designing a sensor model capable of effectively interpreting cross-domain sensor data with minimal complexity. Conventional methods require isolated training for each distinct dataset (domain) due to the inherent personalized characteristics of sensor data within smart environments [47], [71], [74]. Creating a universal sensor model capable of aligning semantic spaces is essential for translating sensor features into human-readable explanations. Additionally, developing a cost-effective strategy for transforming sensor data into vision domain features is crucial to ensure the approach is widely implementable and sustainable, enhancing the accessibility and feasibility of advanced sensor-based applications in smart environments.

B. Contributions

This article introduces several key contributions in sensor data interpretation and translation into the vision domain, thereby enhancing the explainability and applicability of sensor-based activity detection systems. It is also the first article to address this problem in the vision domain. The current method mostly focuses on natural language processing

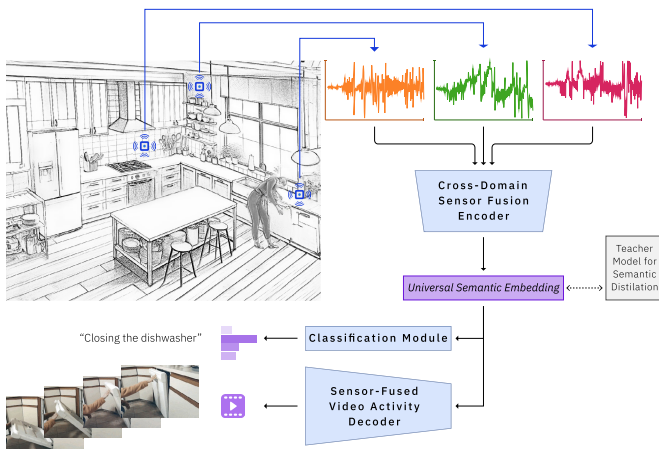


Fig. 1. GeXSe framework: This schematic represents a system for integrating diverse sensor inputs into a cohesive latent space using a cross-domain sensor fusion encoder, which then interfaces with a teacher model to enhance the Universal Semantic Embedding. A classification module interprets the embedded signals, while a sensor-fused video activity decoder translates the embeddings into annotated video activity, such as “closing the dishwasher,” demonstrating the framework’s capability for zero-shot video understanding by leveraging sensor-video correlations.

(NLP)-based solution [5], [34]. Our main contributions are as follows.

- 1) *GeXSe framework*: A novel framework Fig. 1 for intuitively translating sensor data into vision domain features, intelligible to non-experts, enabling multimodal data representation.
- 2) *Multitask encoder*: A multitask encoder coupled with the parallel multibranch multilayer perceptron fast Fourier convolution (PMB-MLP-FFC) model structure, showcasing high performance in smart space activity classification while generating universal representations across various sensor types, facilitating zero-shot downstream tasks by eliminating the need for labeling. Design its multitask loss helps to align the cross-domain sensor data into a semantic representation.
- 3) *Universal decoder*: A pre-trained universal decoder based on stable diffusion for the downstream task of multimodal data representation, reducing computational demands and overall costs by translating semantic representations in a zero-shot fashion.
- 4) *Symbolic reasoning*: Implementation of symbolic reasoning to represent varying sensor data contributions as visual cues within video outputs across different activities and conditions, establishing a low-cost, efficient translation system without necessitating decoder retraining.
- 5) *Cross-domain robustness*: A universal semantic representation enables the model to adapt across diverse domains (various publicly available datasets), underscoring the effectiveness and versatility of generative explanatory sensor system (GeXSe) in real-world applications.

These contributions collectively advance the field by enhancing the interpretability and accessibility of

sensor-based activity detection systems, facilitating intuitive and user-friendly applications in smart spaces.

II. RELATED WORK

A. Activity Sensing in Smart Spaces

1) *Camera-Based HAR*: Camera technology has been pivotal in advancing HAR within the domain of computer vision, supporting a wide array of applications [18], [31], [60], [90]. The declining cost of camera technology enhances its accessibility, though challenges in physical interface accessibility persist [19], [42].

The evolution toward *general-purpose sensing* represents a shift from relying solely on camera-based methods to embracing interconnected, remotely controllable devices. This transition facilitates accessibility, mitigating the barriers associated with camera deployment, including privacy concerns and deployment costs [8], [10], [27]. Consequently, the role of cameras in general-purpose sensing is reevaluated, acknowledging the limitations in capturing certain signals [65].

2) *Emergence of General-Purpose Sensing*: General-purpose sensing embodies the capability to monitor a variety of activities across smart environments through minimal sensor deployment. Modern sensor boards, equipped with a multitude of sensors, offer a versatile foundation for ubiquitous computing applications without necessitating hardware modifications [42]. This approach fosters a distributed sensing paradigm, enhancing the scalability of activity detection and interaction within smart spaces. Unlike traditional sensing methods, general-purpose sensing leverages server-based machine learning analytics to process and interpret the raw data collected, promising enhancements in practical applications through deep learning and explainability [93].

Building on the insights from previous research, it becomes evident that *integrating the strengths of both vision-based and noninvasive sensing modalities remains underexplored*. To bridge this gap, we introduce GeXSe, a system designed to enrich video descriptions by synergistically combining camera-derived insights with data from noninvasive sensors. This approach allows us to leverage the detailed environmental context provided by vision systems alongside the discreet, continuous monitoring capabilities of noninvasive sensors, offering a comprehensive and enhanced depiction of activities within smart spaces.

While general-purpose sensing offers a versatile and scalable approach to activity detection and interaction within smart spaces, the complexity of the machine learning algorithms employed in these systems can hinder their interpretability and transparency. This lack of clarity in the decision-making process of AI-driven systems has led to the emergence of explainable AI (XAI), which aims to address these concerns.

3) *Explainable AI*: The complexity of advanced machine learning methods, particularly in analyzing nonlinear signals, poses significant challenges for nonexperts [75]. This complexity hinders the adoption of deep learning tools in various domains, such as personalized dermatology, where nonspecialists struggle to utilize these tools for diagnostic

purposes [24]. To address these concerns, the field of XAI has emerged, focusing on developing methods to enhance the interpretability and transparency of AI-driven systems [1], [39], [41], [57].

To address these concerns, the field of XAI has gained attention, focusing on making AI-driven systems more understandable [1], [39], [41], [57]. XAI methods aim to clarify how AI models make decisions and can be divided into two main categories: transparent models and post hoc explanation techniques. Transparent models, such as linear regressions and rule-based learning, are easier to interpret but may struggle with complex data relationships [52], [54]. Post hoc methods, on the other hand, provide explanations after the model has made a decision. These include the following.

- 1) Local interpretable model-agnostic explanations (LIME): A technique that creates simple, interpretable models to approximate the behavior of complex models locally around a specific input [71].
- 2) SHapley Additive exPlanations (SHAP): A method that assigns importance values to each input feature based on its contribution to the model's output [47].
- 3) Gradient-weighted class activation mapping (Grad-CAM): A technique that highlights the regions of an input image that are most important for a model's decision [74]. While these post hoc methods provide insights into model decisions, especially in computer vision, they primarily focus on visual explanations and may not capture complex feature interactions.

Notably, GradCAM's application extends to sensor data by converting such data into "image" matrices, adapting to its image-centric analysis framework [5]. This adaptation enables the utilization of GradCAM for multichannel sensor inputs, aligning with its pixel-based processing capabilities. However, the reliance on semantic explanations underscores a broader challenge in delivering comprehensible visual domain explanations, highlighting the ongoing need for more accessible interpretative mechanisms in leveraging sensor data for actionable insights.

This study addresses the challenge of insufficient explanation methods for the visual domain of sensor data. Our objective is to devise an explanation technique finely tuned to sensor data. Our methodology introduces a video-based explanation for nonintrusive sensor data.

4) *Multitask Learning*: Multitask learning (MTL) is a paradigm in machine learning that aims to improve the learning efficiency and prediction accuracy of a model by simultaneously learning multiple related tasks. By sharing representations between related tasks, MTL can leverage commonalities and differences across tasks to generalize better on each task than when learning tasks independently. This approach is particularly beneficial in scenarios where data for some tasks are limited, allowing these tasks to benefit from the knowledge acquired from data-rich tasks. MTL has been successfully applied in various domains, including NLP [17], computer vision [38], and sensor data recognition [56], demonstrating its versatility and effectiveness in enhancing model performance across multiple tasks.

We introduce a novel loss function tailored for MTL, aiming to facilitate the model's ability to acquire a universal representation across diverse domains.

B. Synthetic Vision Domain Description

1) *Visual Representations*: Visual output from sensor data offers several advantages. It captures nonverbal context and subtle details, providing a comprehensive representation of spatial relationships, postures, and interactions within an environment [29]. Moreover, visual representations can transcend language barriers and be universally understood, regardless of linguistic background [40].

2) *Diffusion Model*: One of the most related booming areas is image generation via text. It generate synthesis image from text-based image generation technique [11], [25], [35], [49], [68], [69]. The development of DALL-E2 [55] has revolutionized text-to-image models [64] by introducing a new level of creativity and sophistication. Unlike previous models, which often produced limited or generic images based on textual descriptions, DALL-E2 has the ability to generate unique and imaginative visuals that go beyond what is explicitly stated in the text. This breakthrough has opened up new avenues for creative expression and has the potential to transform industries such as design, advertising, and entertainment. Diffusion model that improves these previous results via diffusion method that beat the state-of-the-art result [16]. Not only texts, sensors that have more real-life "descriptions" also need to derive explainable vision domain features for humans to interact with. The representation of sensor data as an image has been a subject of increasing attention and significant focus [33]. Various methods [23], [45], [79] have been attempted to convert sensor data into an "image," but none of them produce actual images. Instead, these methods are primarily based on NLP and signal graph techniques.

In the realm of generative techniques, diffusion models have garnered significant attention due to their recent achievements. To begin with, a novel approach was presented involving the prediction of noise in image generation. This method produces images from raw Gaussian noises by iteratively refining the noise at each step. Building upon this concept, a plethora of enhancements have been suggested, primarily concentrating on enhancing the quality [73] of generated samples and optimizing the efficiency [77] of the sampling process. The diffusion method has been applied across various domains such as point clouds [48], audio [32], and video [30] generation.

All of these models utilizing diffusion techniques are constructed upon the foundation of the U-Net architecture, with particular emphasis on video generation, recognized as one of the most demanding tasks in the field of machine vision. The prevailing approach for video generation and prediction revolves predominantly around the utilization of the U-Net architecture [9], [30], [83], [88].

3) *Neural Symbolic*: A pivotal aspect of consideration lies in the alignment of purely perception-driven models with the tenets of XAI [66]. Contrary to neural networks, which act as opaque entities incapable of revealing their internal mechanics, symbolic systems stand out for their transparency and intelligibility. For instance, symbolic systems excel not

only in making decisions based on established rules and logic but also in illuminating the rationale behind these decisions. This capability for self-explanation stems from the system's proficiency in delineating and expressing the logical sequence from input to outcome, thus offering insight into its decision-making process. Consequently, these systems can uncover underlying principles, decision-making criteria, and even inherent biases in their reasoning, thereby augmenting the transparency and reliability of AI implementations.

In this study, our objective is to utilize stable diffusion coupled with symbolic reasoning to develop a cost-effective solution for translating sensor data into visual representations. We optimize the decoding step through symbolic reasoning to effectively address the challenges posed by varying sensor types within the dataset. This approach allows for a more versatile and efficient translation process, enabling the seamless conversion of diverse sensor inputs into coherent visual outputs.

III. METHODOLOGY

In this section, we offer an in-depth exposition of GeXSe, beginning with an overview of its design and functionalities. Initially, we delineate the architectural blueprint and core capabilities of GeXSe, setting the stage for a comprehensive analysis. Subsequently, we delve into a rigorous discussion encompassing the theoretical framework underpinning GeXSe, including problem formulation and the application of theoretical principles within the GeXSe system. This approach ensures a holistic understanding of the system's foundation and operational mechanics.

A. GeXSe Overview

GeXSe, depicted in Fig. 1, is a generative system designed to interpret diverse sensor data into unified vision domain explanations, utilizing a modular encoder–decoder architecture. This system allows for the independent training of its encoder and decoder components, enhancing its adaptability and efficiency. Despite their separate training process, the encoder and decoder operate within a unified downstream task structure. Within this framework, the encoder is adept at distilling the essential features of diverse sensor data into a universal semantic representation, facilitating an efficient end-to-end learning process that circumvents the need for additional labeling at the inference stage. The decoder, on the other hand, is finely tuned to translate these encoded representations into vision domain narratives, optimized for zero-shot translation to significantly reduce computational requirements. Unlike traditional approaches, GeXSe integrates multiple sensor channels directly into the video output, providing users with a detailed understanding of the contributions of different sensors in the environment, thereby offering a more comprehensive insight into the dynamics of smart spaces.

B. Sensor Data Knowledge Distillation

The process of cross-domain sensor data fusion tackles the integration of varied sensor datasets from multiple domains, capitalizing on their variety to improve interpretation and prediction across different applications [20], [92].

To illustrate, let us consider a set of users, denoted as $\mathcal{U} = \{u_1, \dots, u_n\}$, where each user u has a corresponding dataset of raw sensor data represented by \mathcal{D}_u . We consolidate these individual datasets into a comprehensive collection, $\mathcal{D}^* = \{\mathcal{D}_{u_1}, \dots, \mathcal{D}_{u_n}\}$ to facilitate a unified analysis. The core objective here is to apply a knowledge distillation strategy, transforming this aggregated sensor data into a more abstract, latent space that captures the underlying patterns and features across all users. This transformation is typically achieved using a teacher–student model framework, designed to harness and transfer knowledge efficiently.

The employed semantic teaching model is XLNet [89], a pre-trained model that leverages permutation-based language modeling. This model efficiently captures bidirectional context, thereby surpassing BERT in sequence order comprehension. Consequently, XLNet facilitates the generation of improved embeddings for nuanced language understanding tasks.

In this framework, the teacher model takes descriptive language annotations as input and generates latent embeddings, $\mathcal{Z}_{\text{teacher}}$, as output. These embeddings encode the semantic content of the activities described by the annotations, serving as a high-level understanding of the sensor data's underlying patterns and structures. Meanwhile, the student model, which is designed to process sensor data, strives to produce latent embeddings $\mathcal{Z}_{\text{student}}$ that closely approximate $\mathcal{Z}_{\text{teacher}}$.

To facilitate this distillation, we segment each user's sensor dataset \mathcal{D}_u into nonoverlapping, fixed-length temporal windows $\mathcal{W}_u = \{w_1, \dots, w_q\}$, each comprising z seconds of consecutive sensor data. The distillation objective can then be articulated as the optimization problem of refining the sensor data within each window w_i into a distilled representation ℓ_i , such that the collection $\mathcal{L}_u = \{\ell_1, \dots, \ell_q\}$ for each user u aligns with the latent embeddings provided by the teacher model. Despite the fundamentally different nature of the sensor data inputs compared to textual annotations, this alignment enables the student model to interpret the sensor data in a manner that is semantically consistent with the language model's understanding of the activities.

Given the inherently personalized nature of sensor data within smart spaces, our methodology facilitates the translation of disparate sensor data types into a unified semantic space. This translation serves as a foundational step toward establishing a universal representation within our system, addressing the challenge of sensor data individualization.

C. Multitask Learning

1) *MTL Definition*: In our framework, MTL is operationalized through a model that concurrently learns to output a representation and perform classification, thereby synergizing two objectives. Specifically, the model is designed to optimize the following dual tasks: first, to distill raw sensor data into a semantically meaningful representation within a latent space, \mathcal{Z} ; and second, to accurately classify the type of activity, $a \in \mathcal{A}$, that generated the sensor data.

Formally, let \mathcal{D} be the dataset containing raw sensor readings, and $\mathcal{T} = \{\tau_{\text{repr}}, \tau_{\text{class}}\}$ represent the set of tasks for representation learning and classification, respectively. The

MTL model, M , is a function that maps an input $x \in \mathcal{D}$ to a tuple $(y_{\text{repr}}, y_{\text{class}})$, where $y_{\text{repr}} \in \mathcal{Z}$ is the latent representation and $y_{\text{class}} \in \mathcal{A}$ is the predicted activity label

$$M : x \mapsto (y_{\text{repr}}, y_{\text{class}}).$$

2) *MTL Loss*: The training objective is to minimize a loss function L that is a composite of the losses for both representation learning, L_{repr} , and classification, L_{class} , effectively enabling the model to learn a shared representation that serves both tasks

$$L = \alpha L_{\text{repr}}(y_{\text{repr}}, y_{\text{true}}) + \beta L_{\text{class}}(y_{\text{class}}, a_{\text{true}}) \quad (1)$$

where α and β are the weighting coefficients that balance the importance of each task during training, y_{true} is the ground-truth latent representation, and a_{true} is the true activity label. This joint optimization encapsulates the essence of MTL in our method, promoting a learning process that is both efficient and robust, particularly suitable for the complex and nuanced domain of smart space sensor data analysis.

3) *Activity Classification via Cross-Entropy Loss*: For the task of multiclass activity classification within our model, we employ the cross-entropy loss function to evaluate the difference between the predicted probability distributions of the model and the actual distribution of activity labels. Given a finite set of k possible activities $\mathcal{A} = \{a_1, \dots, a_k\}$, the model predicts a probability distribution $P(x) = (p_1, \dots, p_k)$ for each input sample x , where each p_i represents the predicted probability of x being associated with activity a_i .

The cross-entropy loss L_{class} for an individual input sample is defined as

$$L_{\text{class}}(P, a_{\text{true}}) = - \sum_{i=1}^k y_i \log(p_i).$$

Here, y is a one-hot encoded vector representing the true activity label, with $y_i = 1$ if the true activity is a_i and $y_j = 0$ for all $j \neq i$, and p_i is the model's predicted probability that the input x belongs to class i .

The objective during the training phase is to minimize the cross-entropy loss L_{class} over all samples in the training dataset. This minimization drives the adjustment of the model's parameters to align the predicted probability distribution more closely with the true distribution of the labels, thereby enhancing the model's precision in identifying the correct activity class among the multiple categories.

4) *Representation Learning via MSE Loss*: To enhance our model's representation learning capability, we employ the mean squared error (MSE) loss function. This approach focuses on minimizing the element-wise squared differences between the predicted latent representation, $Q(z|x)$, and the target latent representation as ground truth, $P(z)$, where z represents the latent embeddings and x is the input sensor data.

The MSE loss, L_{repr} , for an input x is defined as

$$L_{\text{repr}}(P, Q) = \frac{1}{N} \sum_{i=1}^N (P(z_i) - Q(z_i|x))^2.$$

Here, N is the dimensionality of the latent space, z_i is the i th element of the latent vector, $P(z_i)$ corresponds to the i th element of the target latent representation, and $Q(z_i|x)$ is the i th element of the predicted latent representation for the input x . This loss function aims to reduce the discrepancy between the predicted and actual latent representations on an element-wise basis.

Minimizing L_{repr} during the training phase encourages the model to adjust its parameters such that $Q(z|x)$ more accurately approximates $P(z)$ for each element. This is guiding the model toward generating latent representations that are not only meaningful but also semantically aligned with the descriptive nuances conveyed by language. This methodological choice underscores our commitment to achieving a deep semantic synchronization between sensor-derived data and linguistic representations, pivotal for nuanced understanding and analysis within smart environments.

5) *Addressing Activity Representation Challenges Through MTL*: Representing activities in smart environments extends beyond merely mapping the same activities to consistent latent representations. A particular challenge, exemplified by discrepancies in activity labels such as "Open Door 1" versus "Open Door 2," illustrates the need for a model capable of distinguishing between subtly different activities while concurrently merging their representations into a unified latent space.

Challenge:

Differentiate ("open door 1," "open door 2") while

Consolidating into $\mathcal{Z}_{\text{universal}}$.

This scenario underscores the critical role of MTL within our framework. Our model is designed to navigate the dual objectives of identifying distinct tasks and harmonizing their representations simultaneously. The presence of overlapping activities within sensor datasets—where activities may be identical in some perspectives and divergent in others—mandates a solution that not only categorizes these activities accurately but also aligns them within a universal latent space, denoted as $\mathcal{Z}_{\text{universal}}$.

Objective:

Discern ($\mathcal{A}_{\text{identical}}$, $\mathcal{A}_{\text{divergent}}$) and Align within $\mathcal{Z}_{\text{universal}}$.

This dual requirement amplifies the intricacy of our task, compelling the model to not only distinguish between different activities within the same dataset but also to ensure their semantic integration where relevant, as fulfilled by the universal latent space representation, $\mathcal{Z}_{\text{universal}}$. Leveraging MTL, our model adeptly handles these complex objectives, enhancing activity recognition and representation in smart environments.

D. Encoder Model

Fig. 2 (left panel) delineates our encoder's design, trained on three public sensor datasets with 300 epochs and a learning rate of $1e-3$, utilizing the AdamW optimizer. Upon preprocessing, we pass the data through a fast Fourier convolution (FFC) layer for efficient initial feature extraction. The core innovation, the PMB-MLP-FFC module, supports an

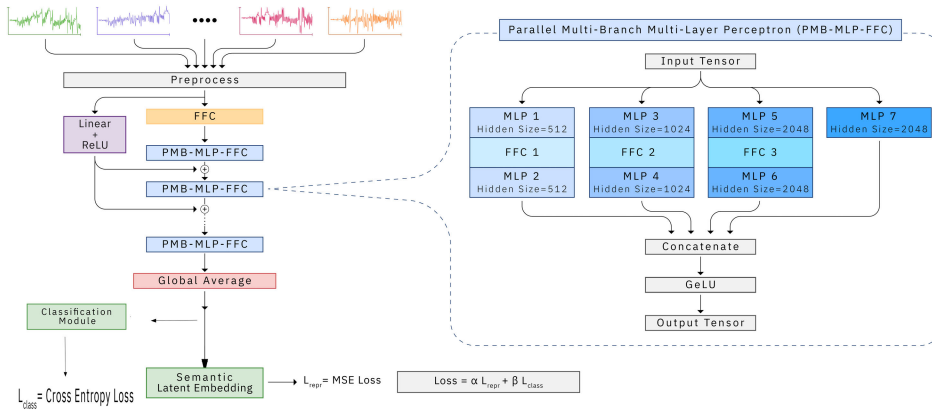


Fig. 2. Encoder model (left) incorporating a novel MTL loss to concurrently optimize for classification and semantic representation. It features a unique PMB-MLP-FFC module (right), adeptly extracting comprehensive features from different sensor types.

arbitrary number (n) of PMB-MLP-FFC blocks in sequence with $n - 1$ residual connections and ReLU activation layers to suit various data complexities while avoiding the vanishing gradients problem. In this study, we implemented two instances of the PMB-MLP-FFC module, linked through a single residual connection, facilitating diverse feature extraction through its multibranch structure with variable hidden layers. Drawing on the strengths of MLP for local feature analysis [46], [82] and FFC for global perspectives [15], our architecture fosters a synergistic local-global feature extraction capabilities—distinctly diverging from conventional inception network frameworks [78]. The architecture comes to a global average pooling layer, which condenses the feature maps into a unified form suitable for subsequent tasks: a classification module powered by MLP and FFC and cross-entropy loss, and a semantic latent embedding module based on a single MLP layer optimized with MSE loss.

1) *Fast Fourier Convolution*: The FFC, introduced by Chi et al. [15], represents a novel approach in machine learning, predominantly applied within computer vision. This study extends its application to smart home sensors, aiming to harness FFC for enhanced feature extraction in HAR. Given the inherent frequency and periodicity in human activities, we capitalized on FFC’s potential for robust feature delineation and irrelevant feature filtration. The pseudocode in Algorithm 1 details the FFC process, converting sensor data into the frequency domain for channel-wise analysis. This conversion facilitates 1-D convolution, normalization, and activation via the ReLU function, allowing pattern learning in Fourier-transformed data. The uniform treatment of multi-channel data in the Fourier domain underscores the versatility of this approach, demonstrating sensor-type agnosticism and signal characteristic discernment.

2) *PMB-MLP-FFC Module*: As elucidated in the previous chapter, the heterogeneity in smart space sensor configurations and data typologies necessitates an architecture adept at accurately interpreting such diverse sensor information.

Convolutional neural networks (CNNs) have been extensively validated as effective for analyzing smart space sensor data, as evidenced by recent studies [5], [63]. Despite their prevalent adoption, CNNs, which rely on convolution as

Algorithm 1 Application of Fast Fourier Convolution on Sensor Data [15]

Input Data: x

```

 $x_{real}, x_{imag} = \text{Real\_FFT}(x)$ 
 $X = \text{Concatenate}([x_{real}, x_{imag}], \text{dim} = \text{channel})$ 
 $X = \text{ReLU}(\text{BatchNorm}(\text{Conv1d}(X)))$ 
 $x_{real}', x_{imag}' = \text{Split}(X), \text{dim} = \text{channel}$ 
 $x = \text{iFFT}(x_{real}', x_{imag}')$ 

```

opposed to the simple matrix multiplications found in *multi-layer perceptrons* (MLPs), introduce additional computational complexity and execution time due to their intricate operational mechanisms [80]. Notably, there have been instances where a singular MLP layer outperformed its CNN counterparts in specific tasks [36].

In response, we introduce the **PMB-MLP-FFC** module (illustrated in Fig. 2, right panel), engineered to capture patterns and extract a wide array of feature types across disparate datasets. This module comprises four distinct parallel branches that process an input tensor with unique contributions while preserving the input dimensionality. These branches are differentiated by their hidden sizes.

Three of these branches each stack an MLP expanding the input tensor’s dimensionality by a factor of two, an FFC, and a subsequent MLP that acts as a bottleneck, shrinking the tensor’s dimension to a quarter of its original size. This design strategy ensures that each branch contributes equally to the output tensor’s dimension. The fourth branch, an MLP-only pathway with the same pattern of a fully connected layer followed by a GELU activation function, similarly contracts its output to a quarter of the input size. Following the individual processing, the output tensors from all branches are concatenated, ensuring the final output tensor matches the input tensor’s size. A GELU activation function is applied to this concatenated tensor to produce the final output.

E. General Purpose Decoder

We introduce a framework designed for sensor-based video explanation, leveraging videos pre-trained across various types

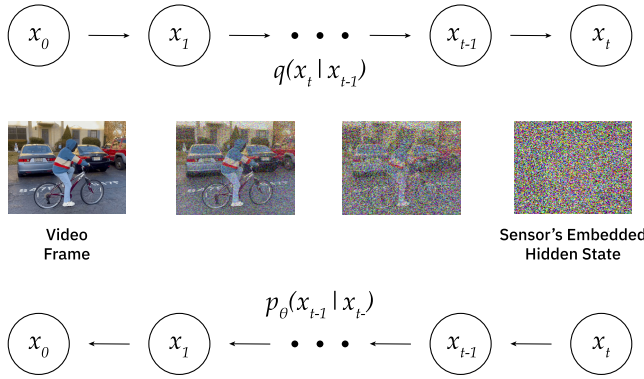


Fig. 3. Sensor-to-video generation process.

and latent representations derived from language models. This setup enables the direct translation of sensor data into video descriptions at a rate of 24 frames/s. Our methodology involves a decoder training strategy that aggregates diverse activity videos, pairing them with corresponding language embeddings. This initial pairing serves to train the decoder, which is subsequently employed to interpret latent representations derived from sensor data across various domains. This approach is specifically tailored to scenarios where computational resources are limited, acknowledging the complexity of training vision domain models like 3-D UNet in comparison to models processing 1-D sensor data.

1) *Stable Diffusion for Activity Description*: Our proposed approach leverages the concept of stable diffusion, mapping sensor data to visual representations through a Markov chain-based process. The framework, as delineated in Fig. 3, introduces a systematic diffusion from the domain of sensor inputs to visual outputs over T discrete steps. This process, represented by $q(x_{1:T}|x_0)$, incrementally infuses noise into the data, thereby facilitating the generation of new latent vectors that gradually conform to a target Gaussian distribution characterized by mean μ_t and variance σ_t^2 , with β serving as a control parameter

$$q(x_{1:T}|x_0) = \prod_{t=1}^T q(x_t|x_{t-1}) \quad (2)$$

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \mu_t = \sqrt{1 - \beta_t}x_{t-1}, \sigma_t^2 = \beta_t I). \quad (3)$$

Crucially, our generative model adapts to the inherent variability of the sensor data via a conditional diffusion model. Equation (4) formalizes this adaptive mechanism, where y represents specific activity conditions detected by the sensors, such as a person running. The encoder first converts these conditions into a latent embedding, which the decoder then utilizes to synthesize a corresponding visual narrative

$$p_\theta(x_{0:T}|y) = p_\theta(x_T) \prod_{t=1}^T p_\theta(x_{t-1}|x_t, y) \quad (4)$$

$$\nabla_{x_t} \log p_\theta(x_t|y) = \nabla_{x_t} \log p_\theta(x_t) + s \cdot \nabla_{x_t} \log p_\theta(y|x_t) \quad (5)$$

$$\hat{\mu}(x_t|y) = \mu_\theta(x_t|y) + s \cdot \sigma_\theta(x_t|y) \nabla_{x_t} \log f_\phi(y|x_t, t). \quad (6)$$

This framework not only simplifies the sensor-to-visual domain translation but also enriches the generated video with nuanced details reflective of the underlying sensor data, offering a sophisticated means to visualize sensor-detected activities.

Practical application example: As depicted in Fig. 1, consider a scenario where sensor arrays in a smart home environment detect the pattern of someone closing a dishwasher. The encoder translates this specific pattern into a latent vector, capturing the dishwasher closing activity's unique characteristics. Subsequently, the decoder, informed by this latent vector and leveraging the stable diffusion process, generates a short video segment that visually depicts the dishwasher closing activity. This showcases the model's ability to transform abstract sensor data into a visually understandable representation, thereby offering a more comprehensive insight into the dynamics of smart spaces.

F. Neural Symbolic Reasoning

Beyond standard video narratives, real-world applications often encounter variations in sensor domain descriptions for identical activities performed by different individuals or in different conditions. By leveraging insights from sensor contributions quantified through sensor channel activations derived from the encoder, we integrate symbolic reasoning to enable nuanced representation of sensor variations corresponding to the same video descriptions [22]. By establishing this low-cost, zero-shot translation system, our approach enables the generation of contextually rich video descriptions adaptable to diverse sensor inputs without necessitating further decoder retraining to accommodate such information.

This process involves the following steps.

- 1) *Quantification of sensor activation*: Sensor data, encompassing a diverse array of physical phenomena, is quantified by the model to produce activation levels for each sensor channel, symbolically represented by a continuous value in the interval $[0, 1]$.
- 2) *Symbolic encoding*: These quantified activations are then symbolically encoded into the video output. Symbolic reasoning is employed to map specific activation levels to predetermined visual cues or modifications within the video, such as altering the speed of video or changing temperature indicator based on the activation of the accelerometer and temperature sensor, respectively.

Formally, let $S = \{s_1, s_2, \dots, s_n\}$ denote the set of sensor channels, and $A = \{a_1, a_2, \dots, a_n\}$ represent the corresponding activation levels, where $a_i \in [0, 1]$ for each sensor channel s_i . The symbolic reasoning process, denoted as $f : A \rightarrow V$, maps these activation levels to a set of visual modifications V , thus transforming quantitative sensor data into qualitative visual representations. This mapping is defined by a set of rules or functions that dictate how activation levels influence specific aspects of the video output, enhancing the interpretability and informativeness of the visualized data

$$V_i = f(a_i) \quad \forall i \in \{1, 2, \dots, n\}. \quad (7)$$

This problem definition underscores our approach to leveraging symbolic reasoning for the intuitive and dynamic

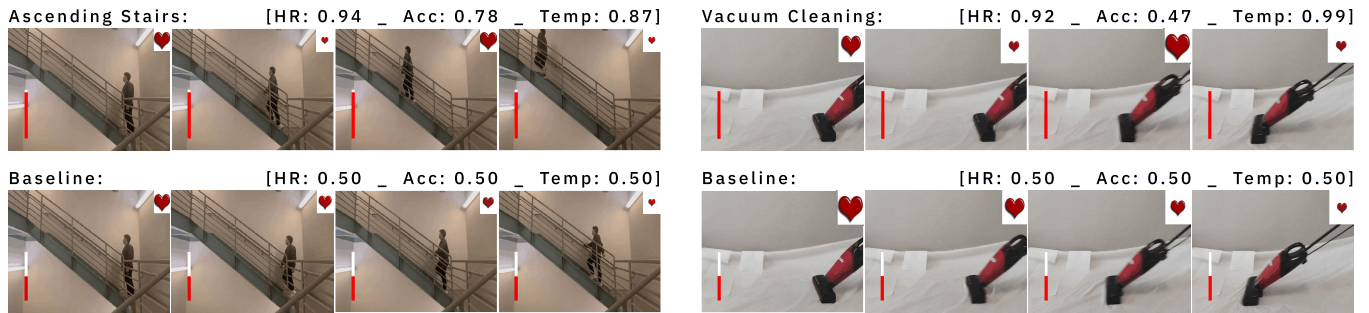


Fig. 4. Four frames of the symbolic reasoning output for “Ascending Stairs” and “Vacuum Cleaning” activities, illustrating variations in sensor activation levels per activity in comparison to their respective baselines. Normalized sensor activation values for heart rate, accelerometer, and temperature are noted. The overall output frames are 24.

representation of sensor data within video outputs, aiming to provide users with a richer understanding of the underlying sensor contributions while obviating the need for retraining the decoder model to offer such information.

Fig. 4 illustrates selected outcomes of the symbolic reasoning for two activities, each with two examples of different activation levels, showcasing the variance in heart rate (indicated by the pulse rate of the heart shape in the top right corner), temperature (represented by a red bar in the bottom left corner), and accelerometer readings (reflected through the manipulated frame rate of video) following symbolic reasoning.

IV. EXPERIMENTS AND RESULTS

In this section, we detail our experimental setup and outcomes, focusing on evaluating the performance of our encoder and the resultant decoder outputs. Initially, we describe the datasets utilized in our study. Subsequently, we assess the encoder’s effectiveness in capturing and representing sensor data. Finally, we appraise the quality of our vision domain explanations through human evaluation across various dimensions.

A. Datasets

In selecting datasets for our experiments, we encountered several challenges that needed to be addressed. The primary issues were the translation of the same activities across different sensor types (using body-mounted sensors, smartphone sensors, and ambient sensors in different datasets) and distinguishing between subtly different activities. To tackle these challenges, we selected three publicly available datasets: PAMAP2, UCI-HAR, and Opportunity. The activity types present in the UCI-HAR dataset can be considered a subset of those in PAMAP2, which helps address the issue of translating the same activity across various sensors. The Opportunity dataset, known for its focus on smart space HAR, addresses the challenge of classifying similar activities differently, such as distinguishing between “Open Door 1” and “Open Door 2.”

The training and testing datasets are constructed from nonoverlapping subjects to ensure generalizability; individuals featured in the testing set were not included in the training phase. The effectiveness of the classification is evaluated using

the macro $F1$ score, serving as the primary metric for testing performance. Below are the details of each dataset.

1) **PAMAP2**: The PAMAP2 [70] dataset is a rich repository of physical activity data, captured from a group of nine individuals (comprising eight males and one female), who were instructed to perform a comprehensive array of 18 lifestyle activities. These activities span from daily household chores, such as lying down, sitting, standing, walking, running, cycling, Nordic walking, ironing, vacuum cleaning, jumping rope, as well as ascending and descending stairs, to various leisure pursuits including watching television, working on a computer, driving, folding laundry, house cleaning, and playing soccer. Participants were outfitted with three inertial measurement units (IMUs) attached to their hand, chest, and ankle, which meticulously recorded data across multiple parameters: accelerometer, gyroscope, magnetometer, temperature, and heart rate, over the duration of 10 h. The constructed PAMAP2 dataset comprises 36 dimensions, with the data segmented into fixed-width sliding windows of 2.56 s, each overlapping by 50%, resulting in 256 readings for every window due to a sampling rate of 100 Hz. These 36 distinct signals formed the input for our model.

2) **UCI-HAR**: The UCI HAR dataset [2] collects activity data from 30 participants using waist-mounted smartphones with inertial sensors, aiming to classify six basic activities: three static postures (standing, sitting, lying) and three dynamic movements (walking, walking downstairs, and walking upstairs). Triaxial linear acceleration and angular velocity data were recorded at a 50-Hz sampling rate, segmented into 2.56-s sliding windows with a 50% overlap, resulting in 128 readings per window. Preprocessing included noise reduction via median and third-order low-pass Butterworth filters with a cutoff frequency of 20 Hz, which also served to segregate the acceleration signal into body motion and gravitational components. This meticulous preparation yielded a dataset with nine processed signal channels, ready for input into deep learning models for comprehensive activity analysis.

3) **Opportunity**: The Opportunity [72] activity recognition dataset is a robust collection of naturalistic activities, acquired in a sensor-rich environment leveraging 72 different environmental and body sensors across 12 subjects, four subjects were released to the public. Utilizing 15 networked sensor systems spanning ten sensor modalities, integrated within the

TABLE I
BENCHMARK HAR MODELS FOR COMPARATIVE ANALYSIS

Model Name	Number of Parameters
vLSTM [53]	120,397
CNN-HAR [87]	310,069
AROMA [62]	515,552
DeepConvLSTM [56]	3,444,205
THAT [43]	6,837,502
UniTS [44]	6,536,475
GeXSe (Ours)	6,392,693

environment, embedded in objects, and worn on the body, the dataset is particularly advantageous for benchmarking a variety of activity recognition methodologies. For our analysis, we focused solely on data from the IMUs, including triaxial accelerometers, gyroscopes, and magnetometers, among others, while omitting quaternion data. This approach provided us with 77 distinct signal channels. The dataset was sampled at 30 Hz, and we employed 3-s windows for data extraction, which yielded 90 samples per window, forming the input for our model.

B. Encoder Performance

This section is dedicated to assessing the classification capabilities of our encoder. Experimental results demonstrate that our design not only facilitates the development of a generalized encoder but also enhances classification performance in comparison to models dedicated solely to classification tasks.

We conducted a comprehensive comparative analysis of various smart space HAR models, highlighting differences in architecture and parameters for a broad evaluation. Table I delineates the models ranging from basic configurations such as CNNs and LSTM networks to more complex models like AROMA and DeepConvLSTM, which have a significantly larger parameter count. Our model features an FFC module for global feature representation, prompting us to include a comparison with a typical transformer-based model (THAT) known for its global feature-capturing capabilities through attention mechanisms. Additionally, we benchmark against UniTS, a Fourier-based machine learning model, which, akin to THAT, aligns with our model in terms of parameter count, enabling a direct comparison to elucidate differences in performance and efficiency. Brief descriptions of these models are provided as follows.

- 1) *vLSTM* [53]: The architecture comprises multiple LSTM cells arranged in a stack, intended for analyzing features within the sensor dataset. This recurrent configuration facilitates the recognition of temporal dynamics and spatial relationships. The performance of this recurrent model in discerning patterns within time series sensor data is demonstrated.
- 2) *CNN-HAR* [87]: The architecture primarily consists of a CNN utilized for recognizing human activities based on data from multiple sensors. To integrate feature maps from diverse sensors, the model formulates layers that extract pertinent features essential for distinguishing various activities. The objective is to synergize feature

learning and classification, thereby enhancing the overall recognition process.

- 3) *AROMA* [62]: It constitutes a multitask learning paradigm employing a residual-based architecture. It simultaneously integrates convolutional, pooling, and fully connected layers as hidden components for multi-task learning.
- 4) *DeepConvLSTM* [56]: It represents a multimodal wearable sensor model founded on LSTM and CNN architectures. This model incorporates a recurrent module designed to capture temporal dynamics and facilitate seamless sensor fusion. In contrast to CNN alone, the inclusion of the recurrent module enables the effective modeling of time-series data's temporal variations.
- 5) *THAT* [43]: It introduces a novel approach to HAR by exploiting WiFi signals' unique variations due to human movement. It addresses the limitations of existing WiFi-based HAR methods by capturing both time-over-channel and channel-over-time features using a two-stream structure. The model incorporates multiscale convolution augmented transformers to detect range-based patterns, demonstrating superior performance and efficiency across real experiment datasets.
- 6) *UniTS* [44]: It integrates short-time Fourier transform (STFT) principles into deep neural networks for sensory time-series classification, eliminating the need for domain-specific feature engineering or intensive hyperparameter tuning. By making STFT weights trainable within its architecture, UniTS adeptly combines time and frequency domain information to discern discriminative patterns in sensory data, demonstrating its effectiveness across various datasets with a notable improvement in classification performance.

As illustrated in Table II, our encoder consistently outperforms the aforementioned models in terms of precision, recall, and $F1$ scores across different datasets. Specifically, in the UCIHAR dataset, which features only six activities, our model outperforms all others by achieving the highest classification accuracy, as well as its own best performance, with an $F1$ score of 0.96. For the PAMAP2 dataset, despite the challenge of classifying a larger variety of activities (12 in total), GeXSe still achieves a notable performance with an $F1$ score of 0.89, indicating its better consistency in performance across diverse activity sets compared to competitors. Lastly, for the Opportunity dataset, known for its imbalance issues, GeXSe again sets the benchmark with the highest $F1$ score of 0.83. This demonstrates its relative robustness and effectiveness in dealing with imbalanced datasets, where it significantly outperforms other models in recognizing activities despite the challenging data conditions.

The lower performance in datasets with a higher number of activities and/or imbalance, like PAMAP2 and Opportunity, suggests that traditional transformers might not be the optimal choice for handling sensor data with these characteristics. Our encoder, with its FFC-MLP design, is specifically engineered to leverage both global and local features effectively, addressing potential issues commonly associated with transformers. This design choice is pivotal in enhancing our

TABLE II
COMPREHENSIVE COMPARISON OF DIFFERENT HAR MODELS ACROSS PAMAP2, UCIHAR, AND OPPORTUNITY DATASETS

Model Name	PAMAP2 Dataset			UCIHAR Dataset			Opportunity Dataset		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
vLSTM [53]	0.86	0.86	0.85	0.92	0.92	0.92	0.70	0.66	0.67
CNN-HAR [87]	0.87	0.87	0.87	0.95	0.95	0.95	0.73	0.71	0.71
AROMA [62]	0.87	0.87	0.87	0.94	0.94	0.94	0.72	0.72	0.72
DeepConvLSTM [56]	0.86	0.87	0.87	0.92	0.92	0.92	0.74	0.73	0.73
THAT [43]	0.87	0.86	0.87	0.87	0.87	0.87	0.72	0.72	0.72
UniTS [44]	0.87	0.89	0.89	0.92	0.92	0.92	0.79	0.78	0.78
GeXSe (Ours)	0.89	0.90	0.89	0.96	0.95	0.96	0.84	0.83	0.83

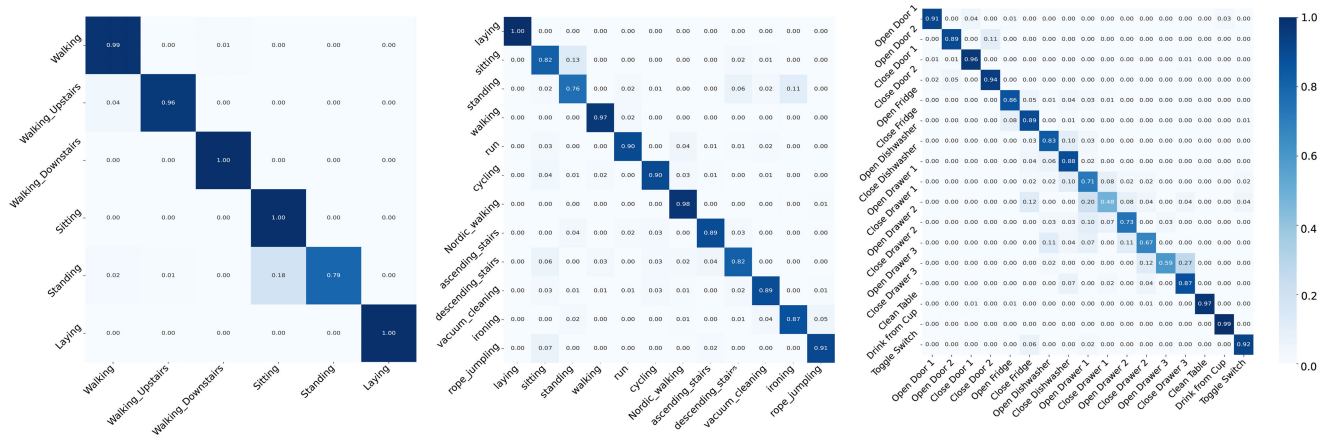


Fig. 5. Normalized confusion matrices: UCIHAR (left), PAMAP2 (middle), and Opportunity (right).

model’s adaptability and performance across datasets with varying characteristics, demonstrating its robustness and the effectiveness of incorporating both global and local feature analysis in activity recognition tasks.

Examining more closely, the PAMAP2 dataset exhibits robust performance across activities, though standing and descending stairs reveal lower recognition rates, as shown in Table III. The UCIHAR dataset, referenced in Table IV, achieves higher recognition rates, albeit with difficulty distinguishing between walking and walking upstairs, indicating the closeness of these activities in the latent embedding space. The Opportunity dataset, mentioned in Table V, highlights a high recognition rate for drinking activities due to a predominance of such data, leading to imbalance. This imbalance impacts intra-class classification, such as differentiating between opening drawer 1 and drawer 2.

The detailed insights from the normalized confusion matrices in Fig. 5 reveal varying intra-class classification performance levels across datasets. The UCIHAR dataset showcases the highest performance, while The PAMAP2 dataset highlights some challenges in differentiating between a few activities like standing and sitting, indicating the closeness of such activities in the latent embedding space. Meanwhile, the Opportunity dataset illustrates more often difficulties in intra-class classification, underlining the model’s struggle with distinguishing closely related activities. Nevertheless, our model’s architecture supports intra-class classification more effectively than the competitors, demonstrating its capability to produce consistent embedded vectors for classification and decoding purposes.

TABLE III
ACTIVITY ANALYSIS OF PAMAP2 DATASET

Activity Name	Precision	Recall	F1-Score
Lying	0.99	1.00	1.00
Sitting	0.84	0.82	0.83
Standing	0.76	0.76	0.76
Walking	0.97	0.97	0.97
Running	0.93	0.90	0.92
Cycling	0.93	0.90	0.92
Nordic walking	0.90	0.98	0.94
Ascending stairs	0.92	0.89	0.90
Descending stairs	0.75	0.82	0.78
Vacuum cleaning	0.90	0.89	0.89
Ironing	0.91	0.87	0.89
Rope jumping	0.76	0.91	0.83

TABLE IV
ACTIVITY ANALYSIS OF UCIHAR DATASET

Activity Name	Precision	Recall	F1
Walking	0.95	0.99	0.97
Walking Upstairs	0.99	0.96	0.97
Walking Downstairs	0.99	1.00	0.99
Sitting	0.83	1.00	0.91
Standing	1.00	0.79	0.88
Laying	1.00	1.00	1.00

C. Decoder Performance

Model interpretation and multimodal data representation are two distinct concepts within the field of data science and machine learning, each addressing different aspects of working with complex data. Model interpretation focuses

TABLE V
ACTIVITY ANALYSIS OF OPPORTUNITY DATASET

Action	Precision	Recall	F1-Score
Open Door 1	0.95	0.91	0.93
Open Door 2	0.94	0.89	0.91
Close Door 1	0.95	0.96	0.95
Close Door 2	0.90	0.94	0.92
Open Fridge	0.89	0.86	0.88
Close Fridge	0.81	0.89	0.85
Open Dishwasher	0.83	0.83	0.83
Close Dishwasher	0.69	0.88	0.77
Open Drawer 1	0.67	0.71	0.69
Close Drawer 1	0.60	0.48	0.53
Open Drawer 2	0.79	0.73	0.76
Close Drawer 2	0.62	0.67	0.64
Open Drawer 3	0.97	0.59	0.73
Close Drawer 3	0.73	0.87	0.79
Clean Table	1.00	0.97	0.98
Drink from Cup	0.99	0.99	0.99
Toggle Switch	0.94	0.92	0.93

on understanding and explaining the decisions, predictions, or outputs generated by a machine learning model. It aims to make the model's operations transparent, allowing humans to comprehend how input data is transformed into outputs, identify the significance of different features, and assess the model's reliability and biases. This is crucial for trust, accountability, and ethical considerations, especially in applications affecting real-world decisions.

On the other hand, multimodal data representation concerns the integration and processing of data from multiple sources or types, such as text, images, audio, and video, within a single model or analytical framework. The challenge here lies in effectively combining diverse data formats to leverage their combined predictive power or insights, which requires sophisticated data preprocessing, feature engineering, and model architecture adjustments. Multimodal representation is essential in applications like automatic content recognition and sentiment analysis, where input comes in various forms and enhances the model's understanding by providing it with a richer, more comprehensive view of the data.

1) *Sensor Channel Activation Feedback Survey*: Our approach employs a multimodal methodology, indicating that the conventional techniques for assessing model interpretability are not applicable to our method. In light of this, we rely on human evaluation to assess the performance of our model, which aims to produce videos where variations in input data rates are perceptibly reflected in the output videos. Hence, we designed two surveys to conduct such human evaluation studies.

In the first survey, we focus on assessing human observers' preferences for different sensor data representation methods within video generation models. This survey aims to compare the effectiveness of our proprietary method against two established model explanation techniques, Gradcam and local interpretable model-agnostic explanations (LIMEs), in communicating the relevance of sensor data to the generated representation.

In the Representation Preference Evaluation Survey, participants are presented with a series of representations of

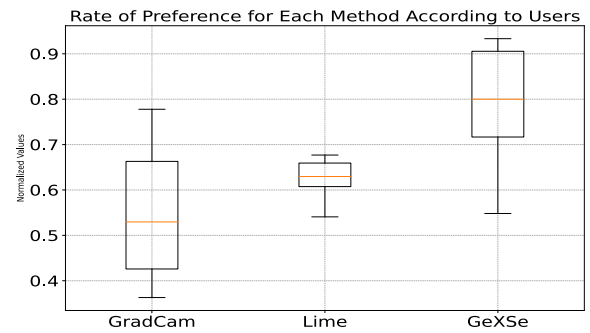


Fig. 6. Human assessment of method preferences.

activities, each annotated with visual highlights generated by different explanation techniques. These highlights aim to indicate areas of the representation that are most influenced by the sensor data inputs. The participants' task is to review these representations and rate how distinguishable and interpretable the activities are represented (from "Strongly Disagree" to "Strongly Agree") in each method. This choice is reflective of the participants' subjective perception of which technique provides the clearest and most understandable visual explanation.

The second survey serves to validate whether changes in acceleration, heartbeat rate, and temperature input data rates are clearly conveyed and discernible in the model-generated videos from a human observer perspective, thus examining the perceived visual influence of the input data on the rendered videos. The premise is that an increase in an input data rate (e.g., acceleration) should result in a perceptible increase in the corresponding visualized measure (e.g., higher frame rate). This principle applies similarly to other data types.

In the Sensor Channel Activation Feedback Survey, participants are presented with a series of videos rendered by our model using varying input rates, along with a baseline video for comparison. Their task is to assess the agreement level (from "Strongly Disagree" to "Strongly Agree") that the variations in input rates are perceptibly reflected in the visualized changes between videos. This allows us to evaluate how understandably our model communicates the relationship between data inputs and video outputs to external viewers.

The results presented below are derived from surveys conducted both in-person and online. We gathered feedback from a total of 60 participants with different backgrounds. To mitigate any potential bias introduced by the order of the survey, we divided participants into separate groups, each comprising 30 individuals, to assess distinct topics. The total valid response is about 900.

2) *Representation Preference Evaluation Survey*: Fig. 6 illustrates the results of the Representation Preference Evaluation Survey. The survey aims to ascertain the preferred methodology for feature extraction from sensor data models. The normalized preference ratings for three distinct methodologies—Gradcam, LIME, and our proposed GeXSe technique—are presented in a box plot format.

From the data, it is evident that GeXSe secures higher median preference ratings in comparison to Gradcam and LIME, indicating a marked preference among users for this

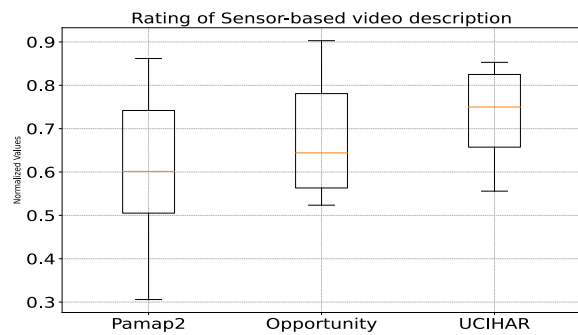


Fig. 7. Human assessment of sensor-based video descriptions.

method. The interquartile range and the whiskers of the box plots reflect the variance in user preferences, with the occurrence of outliers further depicting the diversity of opinions. Notably, the LIME methodology demonstrates a more concentrated distribution of ratings, suggesting a consensus in user perception regarding its interpretability. Conversely, the wider spread in the ratings for GeXSe and Grad-CAM suggests a broader range of user preferences, yet with an inclination toward higher favorability.

Fig. 7 presents the results from the Sensor Channel Activation Feedback Survey. The survey is designed to evaluate the perceived effectiveness of video content modifications triggered by sensor channel activations. Ratings are normalized on a scale from 0 to 1 and presented in a box plot format.

For the PAMAP2 dataset, user consensus is quantified regarding the visual representation of sensor data: the pulse rate depicted by a heart shape for the heartbeat sensor, the level of a red bar for temperature data, and the video frame rate corresponding to accelerometer readings. The Opportunity and UCIHAR datasets, however, focus on representing accelerometer data via frame rate alterations alone. UCIHAR, with its narrower scope of activity types, achieves a higher median rating, suggesting that the representation's simplicity correlates with stronger user consensus. The same pattern is closely followed in the Opportunity dataset case. In contrast, PAMAP2 shows a more dispersed range of user responses, attributed to its variety of sensor types and activity types. This diversity suggests that an increase in the types of sensor and activity data correlates with a wider range of user interpretations and assessments.

V. LIMITATIONS

Beyond the contributions of our study, we aim to outline several limitations pertinent to our research.

- 1) First, the model's capability is currently confined to recognizing a restricted set of activities, which does not fully capture the complexity of real-life scenarios where multiple activities may occur simultaneously. This limitation suggests that the single-activity detection approach may not be sufficient for more intricate situations.
- 2) Second, the inference time for the stable diffusion model, with our settings tuned to $t = 500$, is substantial. To enhance the model's feasibility for real-time or

near-real-time applications, it is imperative to reduce this inference duration.

- 3) Third, the diffusion process at this stage can occasionally produce unstable or inaccurate outputs. This imperfection necessitates attention as the model may render frames that do not precisely represent the input descriptions.
- 4) Lastly, the majority of the datasets used to train and evaluate smart space models are derived from indoor settings. Such datasets lack heterogeneity and complexity present in outdoor or less structured environments. To bridge the gap, there is a pressing need for diverse datasets that more accurately capture the broad spectrum of human activities in various settings.

VI. CONCLUSION AND FUTURE WORK

In this study, we introduced GeXSe, a generative multitask framework for HAR in smart spaces, which proved efficient in fusing sensor data for not only classification but also for generating explanatory visual narratives, accessible to users without expertise in machine learning. By training our encoder to distill cross-domain sensor inputs into a unified semantic representation and leveraging a pre-trained vision domain decoder alongside symbolic reasoning, we effectively manage the variability inherent in sensor data. Our experiments across several public datasets demonstrate the encoder's ability to provide a universal activity representation and achieve superior classification $F1$ scores, with human evaluations further affirming the preferability of our methodology. Future work will explore extending this methodology to more complex real-world scenarios, aiming for faster, practical applications to enhance sensor data comprehension.

REFERENCES

- [1] L. Amgoud and H. Prade, "Using arguments for making and explaining decisions," *Artif. Intell.*, vol. 173, nos. 3–4, pp. 413–436, Mar. 2009.
- [2] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones," in *Proc. Esann*, vol. 3, 2013, p. 3.
- [3] G. Aquino, M. G. F. Costa, and C. F. F. Costa Filho, "Explaining one-dimensional convolutional models in human activity recognition and biometric identification tasks," *Sensors*, vol. 22, no. 15, p. 5644, Jul. 2022.
- [4] A. B. Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.
- [5] L. Arrotta, G. Civitarese, and C. Bettini, "DeXAR: Deep explainable sensor-based activity recognition in smart-home environments," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 6, no. 1, pp. 1–30, Mar. 2022.
- [6] L. Arrotta, G. Civitarese, and C. Bettini, "Knowledge infusion for context-aware sensor-based human activity recognition," in *Proc. IEEE Int. Conf. Smart Comput. (SMARTCOMP)*, Jun. 2022, pp. 1–8.
- [7] A. Bayat, M. Pomplun, and D. A. Tran, "A study on human activity recognition using accelerometer data from smartphones," *Proc. Comput. Sci.*, vol. 34, pp. 450–457, Aug. 2014.
- [8] C. Beckmann, S. Consolvo, and A. LaMarca, "Some assembly required: Supporting end-user sensor installation in domestic ubiquitous computing environments," in *Proc. Int. Conf. Ubiquitous Comput.* New York, NY, USA: Springer, 2004, pp. 107–124.
- [9] A. Blattmann et al., "Align your latents: High-resolution video synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22563–22575.

- [10] M. Boyle, C. Edwards, and S. Greenberg, "The effects of filtered video on awareness and privacy," in *Proc. ACM Conf. Comput. Supported Cooperat. Work*, Dec. 2000, pp. 1–10.
- [11] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," 2018, *arXiv:1809.11096*.
- [12] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Comput. Surv.*, vol. 46, no. 3, pp. 1–33, Jan. 2014.
- [13] L. Chen, J. Hoey, C. D. Nugent, D. J. Cook, and Z. Yu, "Sensor-based activity recognition," *IEEE Trans. Syst., Man, Cybern., C*, vol. 42, no. 6, pp. 790–808, Nov. 2012.
- [14] M. Y. L. Chew and K. Yan, "Enhancing interpretability of data-driven fault detection and diagnosis methodology with maintainability rules in smart building management," *J. Sensors*, vol. 2022, pp. 1–48, Jan. 2022.
- [15] L. Chi, B. Jiang, and Y. Mu, "Fast Fourier convolution," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 4479–4488.
- [16] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 8780–8794.
- [17] G. Durrett and D. Klein, "A joint model for entity analysis: Coreference, typing, and linking," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 477–490, Dec. 2014.
- [18] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly, "Vision-based hand pose estimation: A review," *Comput. Vis. Image Understand.*, vol. 108, nos. 1–2, pp. 52–73, Oct. 2007.
- [19] J. Fails and D. Olsen, "A design tool for camera-based interaction," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, Apr. 2003, pp. 449–456.
- [20] X. Fan et al., "BuildSenSys: Reusing building sensing data for traffic prediction with cross-domain learning," *IEEE Trans. Mobile Comput.*, vol. 20, no. 6, pp. 2154–2171, Jun. 2021.
- [21] W. Fu, K. Breininger, R. Schaffert, N. Ravikumar, and A. Maier, "A divide-and-conquer approach towards understanding deep networks," 2019, *arXiv:1907.06194*.
- [22] A. D. Garcez, M. Gori, L. C. Lamb, L. Serafini, M. Spranger, and S. N. Tran, "Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning," 2019, *arXiv:1905.06088*.
- [23] M. Gochoo, T.-H. Tan, S.-C. Huang, S.-H. Liu, and F. S. Alnajjar, "DCNN-based elderly activity recognition using binary sensors," in *Proc. Int. Conf. Electr. Comput. Technol. Appl. (ICECTA)*, Nov. 2017, pp. 1–5.
- [24] A. Gomolin, E. Netchiporouk, R. Gniadecki, and I. V. Litvinov, "Artificial intelligence applications in dermatology: Where do we stand?" *Frontiers Med.*, vol. 7, p. 100, Mar. 2020.
- [25] I. J. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.
- [26] F. Gu, M.-H. Chung, M. Chignell, S. Valaee, B. Zhou, and X. Liu, "A survey on deep learning for human activity recognition," *ACM Comput. Surv.*, vol. 54, no. 8, pp. 1–34, Oct. 2021.
- [27] A. Guo et al., "VizLens: A robust and interactive screen reader for interfaces in the real world," in *Proc. 29th Annu. Symp. User Interface Softw. Technol.*, Oct. 2016, pp. 651–664.
- [28] S. Ha and S. Choi, "Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 381–388.
- [29] C. E. Hagmann and M. C. Potter, "Ultrafast scene detection and recognition with limited visual information," *Vis. Cognition*, vol. 24, no. 1, pp. 2–14, Jan. 2016.
- [30] J. Ho et al., "Imagen video: High definition video generation with diffusion models," 2022, *arXiv:2210.02303*.
- [31] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Trans. Syst., Man Cybern., C*, vol. 34, no. 3, pp. 334–352, Aug. 2004.
- [32] R. Huang et al., "Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models," 2023, *arXiv:2301.12661*.
- [33] T. Hur, J. Bang, T. Huynh-The, J. Lee, J.-I. Kim, and S. Lee, "Iss2Image: A novel signal-encoding technique for CNN-based human activity recognition," *Sensors*, vol. 18, no. 11, p. 3910, Nov. 2018.
- [34] J. V. Jeyakumar, A. Sarker, L. A. Garcia, and M. Srivastava, "X-CHAR: A concept-based explainable complex human activity recognition model," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 7, no. 1, pp. 1–28, Mar. 2023.
- [35] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8110–8119.
- [36] J. Kaur and A. Kumar, "Speech emotion recognition using CNN, k-NN, MLP and random forest," in *Computer Networks and Inventive Communication Technologies*. Berlin, Germany: Springer, 2021, pp. 499–509.
- [37] B. Khaertdinov and S. Asteriadis, "Explaining, analyzing, and probing representations of self-supervised learning models for sensor-based human activity recognition," 2023, *arXiv:2304.07304*.
- [38] S. Kim, T. G. Purdie, and C. McIntosh, "Cross-task attention network: Improving multi-task learning for medical imaging applications," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, Jan. 2023, pp. 119–128.
- [39] M. A. Köhl, K. Baum, M. Langer, D. Oster, T. Speith, and D. Bohlender, "Explainability as a non-functional requirement," in *Proc. IEEE 27th Int. Requirements Eng. Conf. (RE)*, Sep. 2019, pp. 363–368.
- [40] K. Kveraga and M. Bar, *Scene Vision: Making Sense of What We See*. Cambridge, MA, USA: MIT Press, 2014.
- [41] M. Langer et al., "What do we want from explainable artificial intelligence (XAI)—A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research," *Artif. Intell.*, vol. 296, Jul. 2021, Art. no. 103473.
- [42] G. Laput, Y. Zhang, and C. Harrison, "Synthetic sensors: Towards general-purpose sensing," in *Proc. CHI Conf. Human Factors Comput. Syst.*, May 2017, pp. 3986–3999.
- [43] B. Li, W. Cui, W. Wang, L. Zhang, Z. Chen, and M. Wu, "Two-stream convolution augmented transformer for human activity recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 1, pp. 286–293.
- [44] S. Li, R. R. Chowdhury, J. Shang, R. K. Gupta, and D. Hong, "UniTS: Short-time Fourier inspired neural networks for sensory time series classification," in *Proc. 19th ACM Conf. Embedded Networked Sensor Syst.*, Nov. 2021, pp. 234–247.
- [45] X. Li, Y. Zhang, I. Marsic, A. Sarcevic, and R. S. Burd, "Deep learning for RFID-based activity recognition," in *Proc. 14th ACM Conf. Embedded Netw. Sensor Syst. CD-ROM*, Nov. 2016, pp. 164–175.
- [46] D. Lian, Z. Yu, X. Sun, and S. Gao, "AS-MLP: An axial shifted MLP architecture for vision," 2021, *arXiv:2107.08391*.
- [47] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4765–4774.
- [48] S. Luo and W. Hu, "Diffusion probabilistic models for 3D point cloud generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2836–2844.
- [49] E. Mansimov, E. Parisotto, J. Lei Ba, and R. Salakhutdinov, "Generating images from captions with attention," 2015, *arXiv:1511.02793*.
- [50] J. Marques-Silva, "Logic-based explainability in machine learning," in *Proc. 18th Int. Summer School*, 2023, pp. 24–104.
- [51] H. D. Mehr and H. Polat, "Human activity recognition in smart home with deep learning approach," in *Proc. 7th Int. Istanbul Smart Grids Cities Congr. Fair (ICSG)*, Apr. 2019, pp. 149–153.
- [52] C. Mood, "Logistic regression: Why we cannot do what we think we can do, and what we can do about it," *Eur. Sociol. Rev.*, vol. 26, no. 1, pp. 67–82, Feb. 2010.
- [53] R. Mutegeki and D. S. Han, "A CNN-LSTM approach to human activity recognition," in *Proc. Int. Conf. Artif. Intell. Inf. Commun. (ICAIC)*, Feb. 2020, pp. 362–366.
- [54] N. M. Noh, A. Ahmad, S. A. Halim, and A. M. Ali, "Intelligent tutoring system using rule-based and case-based: A comparison," *Proc.-Social Behav. Sci.*, vol. 67, pp. 454–463, Dec. 2012.
- [55] OpenAI, *Vdalle 2 is a New AI System That Can Create Realistic Images and Art From a Description in Natural Language*, San Francisco, CA, USA, 2022.
- [56] F. J. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.
- [57] A. Páez, "The pragmatic turn in explainable artificial intelligence (XAI)," *Minds Mach.*, vol. 29, no. 3, pp. 441–459, Sep. 2019.
- [58] A. K. Pandey and A. S. Parihar, "A comparative analysis of deep learning based human action recognition algorithms," in *Proc. 14th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT)*, Jul. 2023, pp. 1–7.
- [59] L. R. Parker, P. D. Yoo, T. A. Asyhari, L. Chermak, Y. Jhi, and K. Taha, "DEMISE: Interpretable deep extraction and mutual information selection techniques for IoT intrusion detection," in *Proc. 14th Int. Conf. Availability, Rel. Secur.*, Aug. 2019, pp. 1–10.

- [60] V. I. Pavlovic, R. Sharma, and T. S. Huang, "Visual interpretation of hand gestures for human-computer interaction: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 677–695, Jul. 1997.
- [61] O. A. B. Penatti and M. F. S. Santos, "Human activity recognition from mobile inertial sensors using recurrence plots," 2017, *arXiv:1712.01429*.
- [62] L. Peng, L. Chen, Z. Ye, and Y. Zhang, "AROMA: A deep multi-task learning based simple and complex human activity recognition method using wearable sensors," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 2, pp. 1–16, Jul. 2018.
- [63] S. Pouyanfar et al., "A survey on deep learning: Algorithms, techniques, and applications," *ACM Comput. Surv. (CSUR)*, vol. 51, no. 5, pp. 1–36, 2018.
- [64] A. Ramesh et al., "Zero-shot text-to-image generation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul. 2021, pp. 8821–8831.
- [65] R. Shake, "As the world quieted down in 2020, raspberry shakes listened," Palmira, Boquete, Chiriquí, Panama, Tech. Rep., 2022.
- [66] E. Ratti and M. Graves, "Explainable machine learning practices: Opening another black box for reliable medical AI," *AI Ethics*, vol. 2, no. 4, pp. 801–814, Nov. 2022.
- [67] D. Ravi, C. Wong, B. Lo, and G.-Z. Yang, "A deep learning approach to on-node sensor data analytics for mobile or wearable devices," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 1, pp. 56–64, Jan. 2017.
- [68] A. Razavi, A. V. D. Oord, and O. Vinyals, "Generating diverse high-fidelity images with VQ-VAE-2," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 14866–14876.
- [69] S. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning what and where to draw," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016.
- [70] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," in *Proc. Int. Symp. Wearable Comput. (ISWC)*, 2012, pp. 108–109.
- [71] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1135–1144.
- [72] D. Roggen et al., "Collecting complex activity datasets in highly rich networked sensor environments," in *Proc. 7th Int. Conf. Netw. Sens. Syst.*, 2010, pp. 233–240.
- [73] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 10684–10695.
- [74] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
- [75] D. Shin, "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI," *Int. J. Hum.-Comput. Stud.*, vol. 146, Feb. 2021, Art. no. 102551.
- [76] R. R. Singh et al., "Directive explanations for actionable explainability in machine learning applications," 2021, *arXiv:2102.02671*.
- [77] J. Song, C. Meng, and S. Ermon, "Denosing diffusion implicit models," 2020, *arXiv:2010.02502*.
- [78] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.
- [79] T.-H. Tan, M. Gochoo, S.-C. Huang, Y.-H. Liu, S.-H. Liu, and Y.-F. Huang, "Multi-resident activity recognition in a smart home using RGB activity image and DCNN," *IEEE Sensors J.*, vol. 18, no. 23, pp. 9718–9727, Dec. 2018.
- [80] I. Tolstikhin et al., "MLP-mixer: An all-MLP architecture for vision," in *Proc. 35th Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 24261–24272.
- [81] K. R. Varshney, P. Khanduri, P. Sharma, S. Zhang, and P. K. Varshney, "Why interpretability in machine learning? an answer using distributed detection and data fusion theory," 2018, *arXiv:1806.09710*.
- [82] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017.
- [83] V. Voleti, A. Jolicoeur-Martineau, and C. Pal, "MCVD-masked conditional video diffusion for prediction, generation, and interpolation," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 35, 2022, pp. 23371–23385.
- [84] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognit. Lett.*, vol. 119, pp. 3–11, Mar. 2019.
- [85] Z. Wang, Y. Zhang, Q. Yang, Y. Xu, J. Sun, and S. Liu, "Point cloud quality assessment using 3D saliency maps," in *Proc. IEEE Int. Conf. Vis. Commun. Image Process.*, Jun. 2022, pp. 1–5.
- [86] C. T. Wolf, "Explainability scenarios: Towards scenario-based XAI design," in *Proc. 24th Int. Conf. Intell. User Interfaces*, Mar. 2019, pp. 252–257.
- [87] J. Yang, M. N. Nguyen, P. P. San, X. L. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition," in *Proc. 24th Int. Joint Conf. Artif. Intell.*, vol. 25, Buenos Aires, Argentina, 2015, pp. 3995–4001.
- [88] R. Yang, P. Srivastava, and S. Mandt, "Diffusion probabilistic modeling for video generation," 2022, *arXiv:2203.09481*.
- [89] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 5754–5764.
- [90] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surv. (CSUR)*, vol. 35, no. 4, pp. 399–458, 2003.
- [91] X. Zheng, M. Wang, and J. Ordieres-Meré, "Comparison of data preprocessing approaches for applying deep learning to human activity recognition in the context of Industry 4.0," *Sensors*, vol. 18, no. 7, p. 2146, Jul. 2018.
- [92] Y. Zheng, "Methodologies for cross-domain data fusion: An overview," *IEEE Trans. Big Data*, vol. 1, no. 1, pp. 16–34, Mar. 2015.
- [93] I. Zhou et al., "Internet of Things 2.0: Concepts, applications, and future directions," *IEEE Access*, vol. 9, pp. 70961–71012, 2021.
- [94] W. Zhou, G. Yue, R. Zhang, Y. Qin, and H. Liu, "Reduced-reference quality assessment of point clouds via content-oriented saliency projection," *IEEE Signal Process. Lett.*, vol. 30, pp. 354–358, 2023.