BABEL: A Scalable Pre-trained Model for Multi-Modal Sensing via Expandable Modality Alignment

Shenghong Dai[†]
University of Wisconsin-Madison
Madison, WI, USA
sdai37@wisc.edu

Shiqi Jiang
Microsoft Research
Shanghai, China
shijiang@microsoft.com

Yifan Yang
Microsoft Research
Shanghai, China
yifanyang@microsoft.com

Ting Cao
Microsoft Research
Beijing, China
ticao@microsoft.com

Mo Li
Hong Kong University of Science and Technology
Hong Kong, China
lim@cse.ust.hk

Suman Banerjee
University of Wisconsin-Madison
Madison, WI, USA
suman@cs.wisc.edu

Lili Qiu
Microsoft Research
Shanghai, China
liliqiu@microsoft.com

ABSTRACT

This paper presents BABEL, the expandable modality alignment model, specially designed for multi-modal sensing. While there has been considerable work on multi-modality alignment, they all struggle to effectively incorporate multiple sensing modalities due to the data scarcity constraints. How to utilize multi-modal data with partial pairings in sensing remains an unresolved challenge.

BABEL tackles this challenge by introducing the concept of *expandable modality alignment*. The key idea involves transforming the N-modality alignment into a series of binary-modality alignments. Novel techniques are also proposed to further mitigate data scarcity issue and balance the contribution of the newly incorporated modality with the previously established modality alignment during the expandable alignment process. We provide the comprehensive implementation. In the pre-training phase, BABEL currently aligns 6 sensing modalities, namely Wi-Fi, mmWave, IMU, LiDAR, video, and depth. For the deployment phase, as a foundation model, any single or combination of aligned modalities could be selected from BABEL and applied to downstream tasks.

Evaluation demonstrates BABEL's outstanding performance on eight human activity recognition datasets, compared to a broad range of baselines *e.g.,* the SOTA single-modal sensing networks, multi-modal sensing framework, and multi-modal large language models. BABEL not only improves the performance of individual modality sensing (12% averaged accuracy improvement), but also effectively fuses multiple available modalities (up to 22% accuracy increase). Case studies also highlight emerging application scenarios empowered by BABEL, including cross-modality retrieval (*i.e.,* sensing imaging), and bridging LLMs for sensing comprehension.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Activity recognition and understanding**; • **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**.

KEYWORDS

Multi-modal Sensing, Modality Alignment, Pre-trained Model, Human Activity Recognition

ACM Reference Format:
Shenghong Dai[†], Shiqi Jiang, Yifan Yang, Ting Cao, Mo Li, Suman Banerjee, and Lili Qiu. 2025. BABEL: A Scalable Pre-trained Model for Multi-Modal Sensing via Expandable Modality Alignment. In *The 23rd ACM Conference on Embedded Networked Sensor Systems (SenSys '25), May 6–9, 2025, Irvine, CA, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3715014.3722068

1 INTRODUCTION

Sensing technology, with its distinctive capacity to perceive the physical world, has found widespread application across a multitude of domains, encompassing healthcare, mixed reality, smart driving, and beyond. Over the past several decades, a plethora of sensing modalities have been investigated, each offering a unique and complementary perspective of the physical world. This has led to the emergence of *multi-modal sensing*, an approach that harnesses the simultaneous use of multiple sensing modalities.

Early methods for organizing multiple sensing modalities relied on handcrafted heuristics or features [74], which is proved challenging to scale across various modalities and tasks, due to

[†]Research is done during internship at Microsoft Research.

SenSys '25, May 6–9, 2025, Irvine, CA, USA

ACM ISBN 979-8-4007-1479-5/25/05.
https://doi.org/10.1145/3715014.3722068

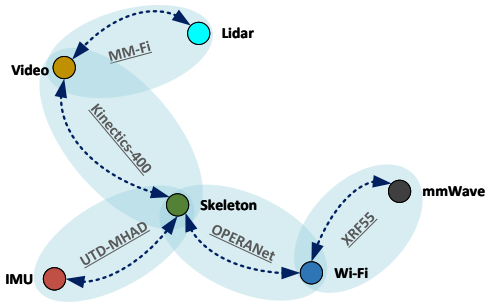


Figure 1: Five public sensing datasets, XRF55 [51], OPERANet [4], MM-Fi [64], UTD-MHAD [9] and Kinectics-400 [24] with binary-paired data cover six modalities.

the complexity of sensing signals and environments. Recent advancements in multi-modal learning have introduced promising solutions [12, 23, 73]. These methods automatically uncover correlations among diverse sensing modalities through supervised or self-supervised learning [11, 22, 27, 40, 44]. Among them, *modality alignment* projects the representations of each sensing modality into a unified and shared space by leveraging paired modality data, demonstrating superior performance [37].

Although modality alignment can effectively organize sensing modalities, existing work [11, 37, 40, 44] is often tailored for specific modalities, necessitating resampling and retraining for different downstream tasks and modality combinations, hindering seamless deployment of sensing applications. Therefore, this paper poses the question: *Can we build a pre-trained multi-modal sensing alignment network as a foundation model*? This model would align common sensing modalities and allow for the integration of new modalities. In the deployment phase, any single or combination of aligned modalities from the model could be selected and applied to downstream tasks directly without retraining.

While modality alignment in AI is a growing research area, its application in sensing presents significant hurdles. The fundamental challenge in supporting multi-modality in sensing is the *data scarcity*, specifically, *(i)* the scarcity of paired data, which is essential for aligning two modalities. For instance, the widely-used CLIP [39] required 400 million image-text pairs for pre-training. In sensing, there lacks paired data from all modalities since some modality data require specialized hardware and expertise to collect. and *(ii)* the scarcity of multi-paired modalities. Existing datasets only contain data from a subset of modalities [4, 9, 24, 64]. For these reasons, existing research [17, 20, 37, 60, 66, 70] struggle to fully incorporate multiple sensing modalities. For instance, due to the limited language-paired data, OneLLM [20] supports a limited number of sensing modalities, *i.e.,* IMU, with subpar performance (see Table 4). Cosmos [37] pioneered the alignment of multiple modalities, but due to the scarcity of multi-paired modalities, it aligns a limited number of modalities *e.g.,* IMU and depth.

In this paper, we present Babel to address this challenge, establishing the first scalable pre-trained network aligning multiple sensing modalities. The design of Babel is underpinned two observations: *(i)* Despite the scarcity of paired data, there exist well-developed encoders or feature extractors for single modality sensing. By leveraging these encoders, the amount of paired data required for modality alignment could be significantly reduced. *(ii)* Even though few datasets provide more than three paired modalities, numerous paired datasets exist that share common sensing modalities. These shared modalities can serve as a bridge for multi-modality alignment (see Fig. 1).

Drawing from these observations, the key idea of Babel is the *expandable multi-modal alignment*, which particularly transforms an N-modality alignment problem into a sequence of binary modality alignments. The *expandability* facilitates the effective utilization of partially paired data in the sensing community. As illustrated in Fig. 1, we could achieve alignment of six modalities through five binary-modality alignments, using the corresponding datasets.

To realize this *expandability*, we introduce three techniques: the pre-trained modality tower (§4), the expandable network architecture (§5), and the adaptive training strategy (§6). Each modality utilizes a modality tower to extract features from raw data. We build these towers using existing singular-modal sensing feature extractors (*e.g.,* LIMU-BERT [59] for IMU), and extend with our alignment modules for aligning with other modality towers; The expandable network architecture enables sequential training phases with only binary-paired samples. Within it we propose the prototype network, shared by all modalities, maintains the knowledge of aligned modalities when adding new ones. Lastly, our adaptive training strategy balances the contribution of newly added modalities to the unified representation, optimally assimilating new knowledge during model growth without disrupting established alignments.

We offer a comprehensive implementation of Babel, including the network architecture, data preparation and processing, as well as the training details. In Babel, we currently align six common sensing modalities: two for wireless sensing, namely Wi-Fi and mmWave, two for mobile sensing, specifically IMU and LiDAR, and two for general vision, namely RGB and depth. As an expandable framework, Babel is allowed for aligning more modalities in the future without retraining aligned modalities. In our work, five datasets are utilized to construct Babel, including UTD-MHAD [9], Kinetics-400 [24], OPERANet [4], XRF55 [51], and MM-Fi [64].

The current pre-trained Babel is evaluated on a typical sensing application, Human Activity Recognition (HAR), across eight datasets, which include both in-domain and out-of-domain datasets [1, 4, 9, 28, 42, 51, 64, 71]. To demonstrate Babel's capability, we compared it to an array of baselines, including the state-of-the-art (SOTA) singular-modal sensing networks [2, 28, 58, 62], multi-modal sensing framework [37], and the emerging Multi-modal Large Language Models (MLLMs) [17, 20, 66, 70].

We use the one-shot learning[1] to evaluate Babel as a foundation model. The evaluation shows that, *(i)* owing to the pre-trained alignment across more sensing modalities and extensive datasets, Babel improve the accuracy of *single modality sensing* by up to 20%, and 12% on average on all six aligned modalities across various datasets. *(ii)* Thanks to the aligned unified embedding space, Babel increases *multi-modal sensing fusion* accuracy by up to 22% compared to current multi-modal frameworks [37]. *(iii)* When comparing with

[1]For each HAR class, we just use only one sample to fine-tune the downstream task classification header.

emerging MLLMs, which are limited in the range of sensing modalities they support, Babel surpasses them by the accuracy of 25.2% across HAR datasets.

Besides HAR, we also present two application studies to highlight Babel's potential as a foundation model. The first is *sensing imaging* to illustrate cross-modality retrieval. With Babel, the original image-to-image diffusion model can be supplemented with non-visual data as input to generate images [41]. The other case aims to bridge the gap between LLM and sensing. By injecting the IMU sensing signal through Babel into the Video-LLaMA [68], the LLM can understand the sensing signals without any retraining.

To summarize, the contributions of the paper include:

- Babel, to the best of our knowledge, is the first expandable foundation model for the multi-modal sensing, currently aligning six sensing modalities.
- Within Babel, we introduce key techniques for learning with scarce paired sensing data and modalities, including the pre-trained modality tower, expandable network architecture, and adaptive training strategy.
- We demonstrate Babel's superior performance compared to a range of baselines. Additionally, we highlight Babel's potential in the field of cross-modality retrieval, and its ability to bridge LLMs for enhanced comprehension of the physical world. Our code is available at aka.ms/babel-project.

2 RELATED WORK

In this section, we introduce the work pertaining to the design of Babel, and highlight our distinctive contribution divergent from the existing research. Specifically, we discuss key advancements in modality alignment, multi-modal sensing, and multi-modal LLMs.

Modality alignment, as an emerging research topic, involving various methods [48, 50]. Of these, contrastive learning (CL) is notable. CL, a self-supervised learning method, differentiates similar and dissimilar samples by comparing positive (similar) and negative (dissimilar) pairs. The goal is to generate representations where similar samples are close, and dissimilar ones are far apart in the feature space. Contrastive Language-Image Pretraining (CLIP) [39] exemplifies the effective use of CL in aligning text and image modalities. CLIP, trained on a large Internet image-caption pairs, learns to associate semantically related texts and images. FOCAL [32] introduces an innovative CL framework with a temporal structural constraint designed for sensing data, addressing challenges in capturing both shared and modality-exclusive features in multimodal time-series data.

Nonetheless, it is still challenging to apply CL to align multiple sensing modalities, due to the data scarcity issue. For instance, CLIP's training necessitates approximately $400,000,000$ image-text pairs, a scale of data that public datasets with paired sensing samples fail to match. In fact, public multi-modal sensing datasets [2, 8, 9], as an example, contain a mere $600-42,000$ sample pairs, a stark contrast to the required volume. Additionally, there exists numerous sensing modalities. The alignment of N sensing modalities generally necessitates a substantial amount of *N-tuple* data. Regrettably, There are no public datasets that cater to the alignment of a greater number of sensing modalities, such as six or more. Babel addresses this fundamental challenge via the proposed *expandable modality alignment* technique.

Multi-modal sensing offers unique abilities to perceive the physical world, incorporates a plethora of methods. For instance, Cosmo [37] pioneered the application of contrastive fusion learning in multi-modal sensing, incorporating RGB, depth and IMU modalities. MESEN [60] employs multi-modal contrastive learning to improve the performance of singular-modal sensing. FM-Fi [53] leverages CLIP through cross-modal contrastive knowledge distillation to improve Radio-Frequency-based human activity recognition with limited labeled data. Nevertheless, these studies are typically crafted for chosen modalities, necessitating retraining for additional ones. In stark contrast to Cosmo, MESEN and FM-Fi, Babel operates as a pre-trained foundational model for multi-modal sensing, facilitating the utilization of any single or multiple aligned modalities for downstream sensing tasks without retraining. Moreover, due to the pre-trained alignment across a broad spectrum of modalities, Babel attains exceptional performance in both single-modal sensing and multi-modal fusion (§8).

The concept of multi-modal sensing is extensively employed across a wide array of applications. For instance, [33] integrates RFID and RGB for recognizing human-object interactions. [52] leverages LiDARs, cameras, and IMU and GNSS devices worn by animals to recognize animal behavior. To locate target individuals, [30] utilizes Wi-Fi Fine Timing Measurements and IMU data to associate individuals in a video with a matched query ID. GaitVibe+ [13] enhances structural vibration-based footstep localization using temporary cameras and vibration sensors for in-home gait analysis. [18] presents an acoustic and camera sensing system that ameliorates range estimation for applications in robotics and others. These applications could benefit through Babel.

Multi-modal LLMs are rapidly evolving to accommodate an increasing number of modalities. Supporting an expanded range of modalities typically necessitates a multi-modal encoder, which projects multi-modal signals into the language embedding space. To construct such an encoder, Meta-Transformer [70] demonstrates the potential of employing a shared transformer encoder across 12 modalities. ImageBind [17] aligns six modalities utilizing solely image-paired data. Similarly, LanguageBind [75, 76] employs the language as the central binding modality to align four modalities. OneLLM [20] aligns eight modalities to language using a singular, unified encoder. CoDi [47] facilitates alignment across language, image, video, and audio modalities.

These studies, however, offer exceedingly limited support for sensing modalities. Indeed, the sole supported sensing modality is the IMU, yet its performance is notably subpar (see the detailed evaluation results in §8). This is primarily due to the high requirement for data with specific modality pairings, such as image-sensing pairs. For instance, ImageBind [17] is only trained on Ego4D [19] for IMU. Their cross-domain capabilities are restricted and they cannot be trained on other sensing modalities without first addressing the issue of data scarcity. In response, we propose crucial techniques for aligning sensing modalities through an expandable architecture, reducing dependence on exhaustive modality pairings. Details are presented in the following.

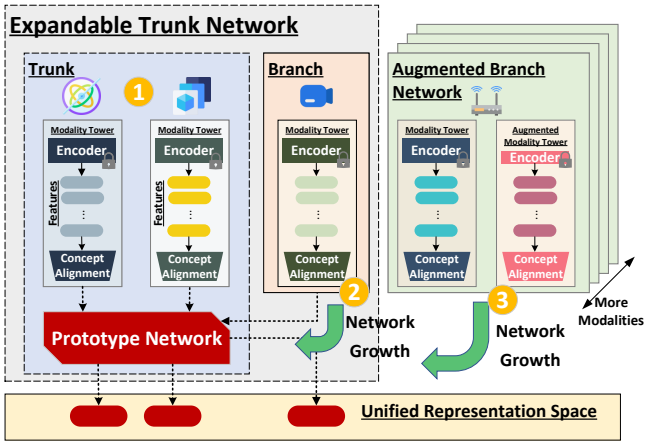


Figure 2: Overview of BABEL.

3 BABEL OVERVIEW

BABEL, to the best of our knowledge, is the first scalable multi-modal pre-trained network, specifically designed for sensing applications, suitable for a multitude of downstream tasks. BABEL consists of the model architecture designs, training strategies as well as the data preparation and processing techniques. In BABEL, we present two designs to build the network with constraint data, namely *pre-trained modality tower* and *expandable model architecture* to cope with the scarcity challenge of paired sensing data and multi-paired sensing modalities.

In the design of the **pre-trained modality tower**, our aim is to harness the power of existing feature extractor within singular modality sensing to construct the modality alignment network, thereby significantly decreasing the necessity for extensive paired training samples.

The crux of this design lies in the efficient alignment of representations across pre-trained encoders. Thereby, we introduce the *modality tower*, consisting of the pre-trained encoder, and the concept alignment module. The encoder could be based on signal processing and neural networks from existing deep learning models. The concept alignment module then aligns embeddings (features) from encoders. During training, pre-trained encoders are frozen, and the concept alignment module is updated.

In the design of the **expandable model architecture**, we try to convert the contrastive training process with *N-tuple* samples into a sequence of training phases involving only paired samples, thereby reducing the need for tupled samples, rendering the alignment of multiple modalities truly feasible.

As illustrated in Fig. 2, we initially align two modalities to form a *trunk* network. We then introduce a new *branch* modality, identifying the *junction* modality within the trunk that pairs with the branch according to available training samples. Through CL, the branch is merged with the trunk to form the updated trunk, by aligning the branch and junction modality. We refer this process as *growth*. The crux of this design lies in effectively maintaining knowledge of aligned modalities while assimilating new insights from the newly merged modality. Thereby, we introduce *prototype network*, which is shared by all modalities and is carefully updated during training with our **adaptive training** strategy.

Adaptive training strategy is explicitly engineered for the sensing modality alignment. Particularly, during each training phase, we aim to create an embedding space where similar sample pairs converge by adjusting each modality's representation. The adjustment weights are vital as modalities contribute differently to the final space. More weight should be given to modalities with more clear signals, while those with more noise or fewer insights should contribute less, to preserve aligned modalities' knowledge. This balance varies depending on the modality combinations, datasets, and tasks. Hence, we propose an adaptive strategy for automatically determining weights.

Next we would introduce these designs in detail.

4 PRE-TRAINED MODALITY TOWER

4.1 Assembling Modality Towers

In the alignment of each modality, our initial step involves constructing a *modality tower*. Subsequent to this, we execute the contrastive learning on these modality towers. The modality tower incorporates two fundamental components: a pre-trained encoder, and a concept alignment module.

Comparing with conventional modality alignment methods *i.e.,* CLIP, BABEL's key design lies in the utilization of a pre-trained encoder within a singular modality, proving particularly effective for sensing modalities. Sensing modalities (e.g., IMU, LiDAR, Wi-Fi) have matured over decades of research, leading to specialized feature extractors. These encoders incorporate domain knowledge in their architecture, training scheme, or signal-processing pipeline. By reusing these encoders rather than designing new ones from scratch, BABEL benefits from the high-quality, domain-specific representations already learned in prior work.

BABEL's effectiveness can be attributed to two key factors. Firstly, the process of assembling the modality tower adheres to the proven method of parameter-efficient fine-tuning (PEFT) [31], a technique notably successful in addressing the vision-language modality alignment problem, as evidenced by models like LiT [67] and APE [43]. The concept alignment module could be regarded as an *adapter* in the context of PEFT practices. Secondly, the successful application of PEFT necessitates that the encoders can capture generic features. For modalities such as vision and language, it typically demands pre-training on a substantial corpus of data, ensuring that the pre-trained model does not exhibit significant domain shift and adequately covers representative features for a majority of downstream tasks.

Pertaining to sensing modalities, the input signals are typically modulated, bearing distinct physical interpretations, thereby making them distinctly defined and explicable in terms of physics. As sensing techniques advance, these representative features are further amplified. As a result, we note that the representative features of sensing modalities for a multitude of downstream tasks often remain consistent. This consistency facilitates our opportunity to leverage singular modality encoders in constructing the modality tower, following the practice of PEFT.

The particular encoder for each modality is chosen based on the following criteria. For modalities dedicated to sensing tasks, such

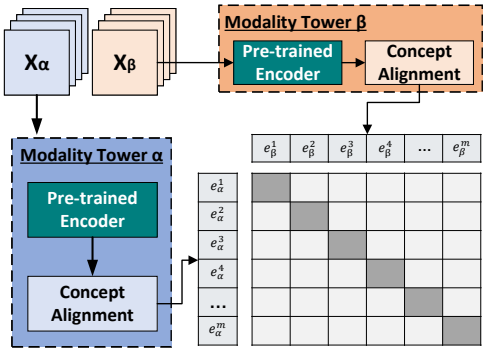


Figure 3: The alignment of two modality towers with pre-trained encoders and concept alignment modules, where e_α^m denotes the m-th embedding of modality α.

as mmWave, we tend to choose signal processing-based encoders, due to their capability to extract universally applicable features with well-defined physical meanings. For more ubiquitous sensing modalities, like Wi-Fi, which are often noisy, we lean towards deep learning (DL)-based encoders, owing to their proficiency in de-noising. We avoid choosing models whose pre-training corpus might exhibit excessive domain shift compared to typical multi-modal sensing tasks. For example, a LiDAR encoder trained solely on autonomous-driving road scenes might not generalize well to indoor human-tracking data. Conversely, encoders trained on more general sets (e.g., broad 3D shapes for LiDAR, a diverse set of action videos) yield a feature space that better suits the variety of tasks Babel must handle. For modalities with large variation or especially noisy signals (e.g., Wi-Fi), relying on a single pre-trained encoder can be limiting. In such cases, we introduce modality tower augmentation, where multiple encoders are employed for the same modality, as elaborated in §4.3. Eventually we evaluate and compare the selected candidates by fine-tuning and testing them on a variety of singular modality datasets. The encoder demonstrating superior generality is chosen. To align modality embeddings, we employ MLP layers as the concept alignment module. MLP-based projection layers have been widely used in contrastive learning frameworks [10].

4.2 Aligning Modality Towers

Upon assembling the modality tower for a given modality, we strive to align them through the contrastive learning. Next, we would illustrate our modality alignment process using the alignment of two modalities as an example. The alignment of multiple modalities would be discussed in §5.

As illustrated in Fig. 3, given the dataset $E_{\alpha\beta}$ comprising paired samples of modality α and modality β, our first step is to structure the positive pairs P and negative pairs Z essential for the contrastive learning process. Specifically, the dataset $E_{\alpha\beta}$ includes sample pairs $(\chi_\alpha, \chi_\beta)$ that are initially synchronized. For instance, in the UTD-MHAD dataset [9], Each sample pair signifies a sequence of IMU readings and a concurrent video recording series of the same human activity, captured within a span of 5 seconds. From the dataset $E_{\alpha\beta}$, we randomly select a batch M comprising m sample pairs. Within this batch, for a given sample of modality α, denoted as χ_α^i where $i \in N$, we construct its corresponding positive pair P_α^i and negative pairs Z_α^i in the following manner,

$$P_\alpha^i = (\chi_\alpha^i, \chi_\beta^i), 1 \le i \le m, \tag{1}$$

$$Z_\alpha^i = \{(\chi_\alpha^i, \chi_\beta^j)\}, 1 \le i, j \le m, i \ne j, \tag{2}$$

Likewise, we can construct the positive pair P_β^i and negative pairs Z_β^i for the ith sample of modality β within the batch M. Ultimately, for the batch M consisting of m pairs, we could derive m positive pairs and $m^2 - m$ negative pairs, which will be utilized in the sequential contrastive learning.

Throughout the training phase, the assembled positive pairs P and negative pairs Z are processed through the modality tower. The contrastive loss L is computed on a per-batch basis for each batch M,

$$L_{\alpha\beta}^M = \frac{L_{\alpha\leftarrow\beta}^M + L_{\beta\leftarrow\alpha}^M}{2}, \tag{3}$$

where $L_{\alpha\leftarrow\beta}^M$ and $L_{\beta\leftarrow\alpha}^M$ denote the computed contrastive loss transitioning from modality β to modality α and vice versa within the batch M, as defined subsequently,

$$L_{\alpha\leftarrow\beta}^M = -\sum_{i=1}^{m} \log\left(\frac{\exp(\text{sim}(P_\alpha^i)/\tau)}{\sum \exp(\text{sim}(N_\alpha^i)/\tau)}\right), \tag{4}$$

where τ is a temperature parameter employed to scale the logits. In our implementation, we set τ to 0.07. The function sim represents the cosine similarity function utilized to examine the output embeddings from Γ_α and Γ_β. Similarly, we can compute $L_{\beta\to\alpha}^M$. Eventually we use $L_{\alpha\beta}^M$ to update the concept alignment modules of modality towers of α and β.

As a pre-trained network, when Babel is incorporated into downstream tasks, we would introduce an additional task-specific network. For instance, a classifier head is introduced for activity classification tasks. Owing to the modality alignment, the aligned embedding from each modality can be straightforwardly concatenated for downstream tasks. As will be demonstrated in the evaluation, the output embeddings, enhanced by modality alignment, are significantly superior. Consequently, we can attain SOTA results even with a very simple classifier, such as a 2-layer MLP, when only applying one-shot learning.

4.3 Augmenting Modality Towers

We also propose to augment the modality towers by employing multiple encoders for the particular modality. The concept of modality tower augmentation is inspired by model ensembling [5, 7], where multiple weak learners combine to create a stronger one, improving accuracy and performance. This method has proven to effectively decrease variance and bias in each weak learner.

In Babel, we would construct an augmented modality tower when incorporating additional encoder. We align the augmented modality towers in accordance with the process delineated in 4.2. Specifically, We construct two modality towers, Γ_α^ϵ and Γ_α^η, using pre-trained encoders ϵ and η respectively. We align these towers using positive pairs $P_\alpha^i = (\chi_\alpha^i, \chi_\alpha^i)$ and negative pairs $Z_\alpha^i = \{(\chi_\alpha^i, \chi_\alpha^j)\}$

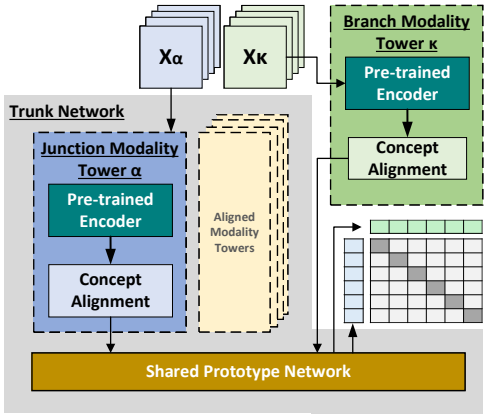


Figure 4: The alignment of multiple modalities with the prototype network in the expandable network architecture.

where $i \neq j$. The alignment is achieved through loss functions from Equations 3 and 4. The similarity sim is computed using output embeddings from both towers.

5 EXPANDABLE MODEL ARCHITECTURE

5.1 Prototype Network

Aligning multiple sensing modalities (such as six or more) with partially paired modalities is challenging. In response to this, one of key designs in Babel is the expandable model architecture, which transforms the training process for N modality alignment into a series of two modality alignment phases, exploiting existing datasets with paired modalities.

To elaborate, consider the alignment of three modalities: α, β, κ, with the available datasets $E_{\alpha\beta}$ and $E_{\alpha\kappa}$. We initially employ $E_{\alpha\beta}$ to align the modalities α and β, as discussed in §4.2, yielding the network $H_{\alpha\beta}$, which we term the *trunk* network. Subsequently, we aim to integrate an additional modality κ into the trunk $H_{\alpha\beta}$.

Given that dataset $E_{\alpha\kappa}$ provides corresponding pairs between the modalities α and κ, we designate α as the *junction* modality. From the trunk $H_{\alpha\beta}$, we select the trained modality tower Γ_α. We then construct a new modality tower Γ_κ, referred to as the *branch*. This branch is integrated into the trunk network by aligning the junction modality tower Γ_α with the branch modality tower Γ_κ, utilizing samples from the dataset $E_{\alpha\kappa}$. We refer to this procedure as network *growth*. Fig. 4 illustrates the network growth in our expandable network architecture.

The challenge of facilitating network growth lies in maintaining the knowledge of previously aligned modalities while concurrently assimilating new insights from the additional modality. Therefore, during this growth phase, it is not suitable to directly align Γ_α and Γ_κ as outlined in §4.2, since any updates to the junction modality Γ_α may significantly disrupt the already aligned modalities, such as modality β.

To this end, we introduce the prototype network. As shown in Fig. 4, it is specifically incorporated into the trunk network, succeeding the concept alignment module of each modality tower. The prototype network is shared across all modality towers within the trunk network. It serves as a coordinating entity for all the learned knowledge across aligned modalities. Therefore by adjusting the updates on the prototype network, we could strike a balance between acquiring new knowledge from the branch modality and avoiding catastrophic forgetting of the trunk network.

Revisiting our previous example, during the initial alignment of modalities α and β, we concurrently update the prototype network Υ while training the concept alignment module of the modality towers Γ_α and Γ_β. Subsequently, during the network growth phase involving the branch modality κ and the junction modality α, the contrastive learning process would updates the branch and junction modality tower Γ_κ along with the prototype network Υ.

In our implementation, the structure of the prototype network is kept relatively straightforward, resembling a 2-4 layer MLP. Despite its simplicity, this design enables several advantages for the alignment of multiple modalities. First, during each network growth phase, it allows us to utilize different datasets, even for disparate tasks. Second, this design facilitates the repeated enhancement of aligned modalities using varied datasets. By assimilating insights from these different datasets, it becomes feasible to construct a pre-trained network with substantial generality.

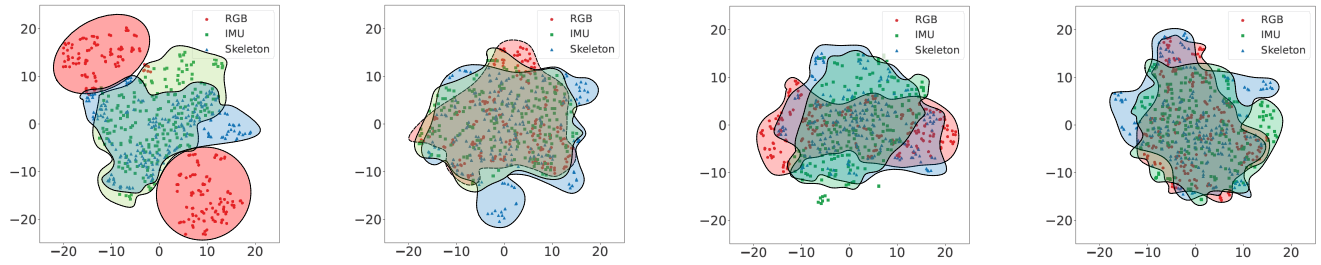
Together with the prototype network, we also devise the adaptive training strategy to regulate the extent to which the trunk network acquires new knowledge, which would be discussed in §6.

5.2 Growth Orders

Babel transforms the N-tuple modality alignment into a sequence of two-modality alignment phases, thereby raising a potential question regarding the differences between the conventional completed alignment and our expandable alignment with varying modality growth orders. The insight of our prototype network is that it maintains a shared set of parameters across all modalities, inherently encoding common features learned from previous alignment phases. When aligning a new branch modality with the junction, the prototype network is updated in a way that partially shifts the shared embedding space, but does not overwrite it completely. This mitigates catastrophic forgetting [16], a prominent challenge in continual learning, which is usually addressed by utilizing shared representations to preserve previously learned information [36, 38].

To analyze this, we take a three-modality alignment, *i.e.,* IMU, skeleton, and video from UTD-MHAD dataset [9], as an example. As depicted in Fig.5, we utilize t-SNE to render the representation space of each modality visible. As evident in Figure 5a, before alignment, features that have not undergone alignment training exhibit significant distribution differences. Fig. 5b shows the conventional triplet alignment successfully bridge the modality gaps, aligning the three modalities. In contrast, the expandable network architecture within Babel employs a sequence of two-modality alignment training phases as a replacement for the joint alignment. As illustrated in Fig. 5c, we initially align the IMU and skeleton modalities followed by the video modality, effectively bridging the modality gaps as well.

Our method is flexible regarding alignment order. Fig. 5d shows representations from each modality achieved by an alternately ordered network: first aligning skeleton and video, then IMU. Despite



a Without alignment **b Triplet alignment** **c BABEL (IMU-Skeleton-Video)** **d BABEL (Skeleton-Video-IMU)**

Figure 5: t-SNE representations of three modalities obtained by different modality alignment approaches.

varying sequences, a common representation space is achievable. Further evaluation would be discussed in Section 8.1.5.

6 ADAPTIVE TRAINING STRATEGY

We further propose our training strategies to optimally integrate the insights derived from the newly aligned modality during network growth. Specifically, we implement two strategies for the training of the concept alignment module and the prototype network, respectively.

For the training of the concept alignment module during network growth, we employ adaptive weighted contrastive training. The key of this design lies in dynamically adjust the proportion of proximity between modalities during the modal alignment process.

As per Equation 5, the contrastive loss in aligning modality α and β includes two parts: $L_{\alpha\leftarrow\beta}$, the loss when β approximates α, and $L_{\beta\leftarrow\alpha}$, the loss when α approximates β. We find reliable and unreliable modalities in various modality combinations and datasets. Naturally, modalities with robust encoders and abundant data are more reliable, so we expect less reliable ones to converge towards them. During network growth, careful updates are needed in the junction modality tower to add insights from the branch without disrupting aligned modalities. Hence, we integrate weights into Equation 3 as follows:

$$L^M_{\alpha\beta} = \frac{w_{\alpha\leftarrow\beta}\cdot L^M_{\alpha\leftarrow\beta} + w_{\beta\leftarrow\alpha}\cdot L^M_{\beta\leftarrow\alpha}}{2}, \tag{5}$$

where M represents a batch randomly drawn from the dataset $E_{\alpha\beta}$, and $w_{\alpha\leftarrow\beta}$ and $w_{\beta\leftarrow\alpha}$ denote the normalized weights. Intuitively, we lean towards attributing a larger weight $w_{\alpha\leftarrow\beta}$ if modality α is deemed more reliable and established, while a smaller weight is assigned otherwise.

Identifying the appropriate weights presents a challenge. A static weighting scheme is suboptimal as each modality may differ in respect to data volume and quality, encoder proficiency, as well as the fresh insights and contributions it brings to the aligned modalities. As such, we opt for a dynamic weighting strategy. Particularly, we employ gradients as an indicator to adaptively modify the weights,

$$w^M_{\alpha\leftarrow\beta} = \frac{1}{\|\nabla^M_{\alpha\leftarrow\beta}(\Gamma_\alpha, \Gamma_\beta)\|}, \tag{6}$$

where ∇ represents the accumulated gradients of all parameters within the concept alignment modules of the modality towers Γ_α and Γ_β when computing the loss $L^M_{\alpha\leftarrow\beta}$ within the batch M. We calculate $w^M_{\beta\leftarrow\alpha}$ in a similar way. Then we normalize them as,

$$w^M_{\alpha\leftarrow\beta} + w^M_{\beta\leftarrow\alpha} = 1, \tag{7}$$

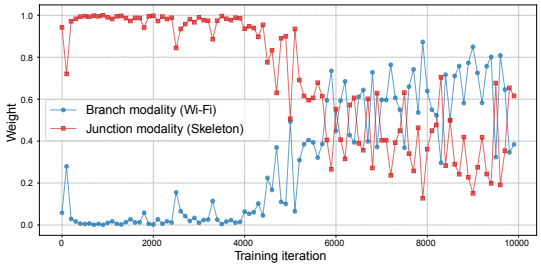


Figure 6: Adaptive training weights of branch and junction modality during the network growth.

Gradient magnitudes effectively indicate how each modality contributes to the alignment process. During network growth, small gradients in the junction modality tower prompts a higher weight, bringing the branch network nearer to the trunk. When the branch modality tower's gradients are significant, the assigned weight speeds up the absorption of insights from the trunk network, ensuring alignment in the unified representation space. Our approach employs bidirectional contrastive learning for each pairwise alignment phase (e.g., $\alpha \leftarrow \beta$ and $\beta \leftarrow \alpha$), but rather than maintaining fixed equal weights between these directions—which can lead to either insufficient adaptation of the branch modality or excessive perturbation of the trunk network—we dynamically adjust the update magnitudes for both modules by monitoring their gradient norms during training. When the branch exhibits larger gradients, indicating that the branch modality is farther from alignment, we increase the weight on the junction modality (pulling the new modality "in"). Conversely, if the junction modality's gradient is large, this suggests that the new modality is relatively close, allowing the junction side to shift more than at the beginning. This adaptive weighting scheme naturally preserves previously learned representations and, through sufficient training iterations, ensures convergence to a consistent common representation, regardless of the order in which modalities are incorporated.

Table 1: Datasets and their corresponding data pairs utilized to train the hexa-modal alignment network.

Dataset	Modalities	# Train Pairs	# Test Pairs
UTD-MHAD [9]	IMU and Skeleton	613	248
MM-Fi [64]	LiDAR and Video	17,528	3,132
OPERANet [4]	Wi-Fi and Skeleton	25,433	5,086
XRF55 [51]	mmWave and Wi-Fi	30,000	12,900
Kinetics-400 [24]	Video and Skeleton	234,619	19,761

Fig. 6 shows the dynamic weight adaptation in the multi-modal alignment network construction using Babel. This merges Wi-Fi as a branch modality into the trunk network, with the skeleton as the junction modality, using the OPERANet dataset for training [4]. Initially, the skeleton modality, enriched with trunk network's aligned knowledge, is more reliable than the Wi-Fi branch modality, thus, it's assigned a near-one weight to speed up convergence with the junction modality. After around 6,000 training iterations, alignment is essentially achieved. Then, our dynamic weight adaptation mechanism adjusts to enable knowledge exchange between the junction and branch modalities, creating a comprehensive representation space.

For the training of the prototype network during network growth, we employ the exponential moving average (EMA) methodology. This strategy aids in preserving stability in the prototype representations by slowly incorporating fresh information while safeguarding the accumulated knowledge. We supplement this with knowledge distillation during the EMA process. This technique assists in preserving crucial information gleaned from prior modalities whilst incorporating novel ones.

7 IMPLEMENTATIONS

7.1 Data Preparation

Overall, we utilize five datasets' training sets for the alignment, as itemized in Table 1, comprising paired samples across divergent dual modalities. These datasets are for human activity recognition (HAR) tasks, but the certain activities are totally different. Despite the provision of activity labels within these datasets, we adopt a self-supervised training approach, labels are not used. Throughout each dataset, we split into train and test set, detailed in Table 1.

Skeleton[2] and IMU pairs. UTD-MHAD dataset [9] is used, encompassing the skeleton and 9-axis IMU data pairs, captured via the Microsoft Kinect sensor and the wearable inertial sensor with respective sampling rates of 30Hz and 50Hz. The dataset embodies 27 distinct actions performed by 8 subjects. Each subject repeated the action for 4 times, totaling 861 paired samples. We use 613 pairs for the training.

LiDAR and video pairs. MM-Fi dataset [64] is used, which contains 27 distinct actions performed by 40 human subjects. The LiDAR is collected in the point cloud format. MM-Fi [64] dataset provides 17,528 pairs for our training.

Wi-Fi and skeleton pairs. OPERANet dataset [4] is used, which contains the paired Wi-Fi CSI and skeleton data. The Wi-Fi CSI is gathered from the Intel 5300 platform across 30 subcarriers, employing a sampling rate of 1600Hz, with 3 transmitters and 3 receivers. The skeleton data is obtained from the Microsoft Kinect sensor. The dataset encompasses roughly 8 hours of annotated measurements collected in two different rooms with 6 participants performing 6 daily activities. OPERANet provides 25,433 pairs for our training.

mmWave and Wi-Fi pairs. XRF55 dataset [51] is used, which is collected from a TI IWR6843ISK radar for mmWave and Intel 5300 for Wi-Fi CSI. It includes HAR data from 39 subjects performing 55 unique actions, each repeated 20 times. In total, 30,000 pairs are provided for training.

Video and skeleton pairs. Kinetics-400 dataset [24] is used, which contains 400 distinct human action classes, each characterized by at least 400 video clips extracted from YouTube. Each clip, approximately 10 seconds long, portrays a variety of human actions. The skeleton is extracted from use clips using OpenPose [6]. Overall, as a dataset in vision modality, the Kinetics dataset provides 234,619 training pairs.

7.2 Data Augmentation

We implement two data augmentation techniques on the raw UTD-MHAD training data, ultimately enlarging the data pairs by 600×. *(i)* Down-sampling. Raw pairs undergo down-sampling at different ratios, simulating diverse sampling rates on various devices or accelerating the action at distinct ratios. This method augments the raw pairs by a factor of 300×. *(ii)* Action-segmentation. The raw action sequence is randomly truncated, simulating incomplete activity sensing. We ensure the segmented sequence's shortest length is over 50% of the original length. This method amplifies the raw pairs by a factor of 300×.

7.3 Pre-trained Encoders and Concept Alignment Architecture

Next we introduce the pre-trained encoders we use for building the modality alignment network, along with the concept alignment architecture.

IMU. We utilize the LIMU-BERT encoder [59], renowned for its proficiency in generating generalized representations. It is pre-trained on a range of IMU datasets.

Skeleton. We utilize the Spatial-Temporal Graph Convolutional Network (ST-GCN) [61] as our encoder, which is pre-trained on extensive datasets, notably the NTU-RGBD [45].

Video. We employ ResNet3D model [49] as the encoder, which is pre-trained on Kinetics-400 dataset [24].

Wi-Fi. For Wi-Fi CSI, we fail to obtain one powerful pre-trained encoder. Therefore we apply multiple encoders to augment the modality tower of Wi-Fi. Specifically, we utilized a Vision Transformer (ViT) and a combination of Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU) as our encoders. They are pre-trained on UT-HAR [65] datasets.

mmWave. We employ the signal processing based encoder for this modality. We use doppler fast fourier transform (FFT) and angle FFT, generating range-doppler heatmaps and range-angle heatmaps, respectively. We supply an additional spatial ResNet18 [51] to further extract features from them.

[2]Depth signals undergo a conversion into a human skeleton format. As such, we employ the term *skeleton* to denote the depth modality

LiDAR. We use the Point Transformer [72], which is pre-trained on the ModelNet40 dataset [55]. The encoder cannot extract temporal features, we add an additional ST-GCN as the additional temporal feature extractor, which is pre-trained on the NTU-RGBD [45] dataset.

The **concept alignment architecture** comprises two Multi-Layer Perceptrons (MLPs) for each modality. This concept alignment module is customized for each modality, with the input dimensionalities determined by their encoders: LiDAR (60), Skeleton (60), IMU (72), Video (512), Wi-Fi (900), and mmWave (1024).

7.4 Training Details

We commence the training process with the IMU and skeleton modalities. Subsequently, we integrate the video modality, aligning it with the pre-existing skeleton modality. Next, we incorporate the Wi-Fi modality into our framework, leveraging the paired Wi-Fi and skeleton data. This is followed by the introduction of the mmWave modality, which is linked with the intermediate Wi-Fi modality. Ultimately, we incorporate the LiDAR modality, capitalizing on its integration with the paired video modality.

We employ the AdamW optimizer [35] with a batch size of 256 and an initial learning rate of 1×10^{-4}. Given a batch size of $m = 256$, we construct 256 positive pairs and 65,280 negative pairs for contrastive learning, following the construction method detailed in Section 4.2. For each phase of network growth, we judiciously allocate a varying number of training epochs, typically up to 500, or cease the training process once convergence is attained. The learning rate for downstream tasks is adjusted between 0.001 and 0.1. We train on two NVIDIA A100 GPUs, spending around 20 hours to align six modalities.

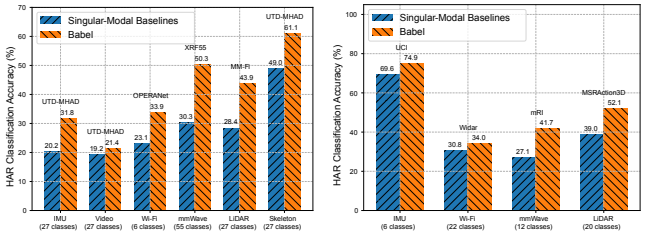
8 EVALUATION

We evaluate pre-trained Babel by employing a typical downstream sensing task, human activity recognition (HAR). Furthermore, we demonstrate two applications enabled by Babel: cross-modality retrieval and LLM integration.

8.1 Evaluation on HAR

We evaluate Babel on 8 datasets, comprising 4 in-domain datasets including UTD-MHAD [9], OPERANet [4], XRF55 [51] and MM-Fi [64]. We utilize the test pairs outlined in Table 1. Additionally, we assess on 4 out-of-domain datasets, which were not part of the pre-training datasets at all: UCI [42], Widar3.0 [71], mRI [1] and MSRAction3D [28]. Out-of-domain evaluation is critical for understanding model generalization in real-world sensing applications. In practical deployments, sensing models frequently encounter new environments and previously unseen subjects or activities. By evaluating on out-of-domain datasets, we can measure how well Babel adapts to new scenarios without any additional training.

The results for Babel are obtained in a *one-shot* setting, which serves as a widely used benchmark for evaluating the generalization capability of a backbone model [14, 25]. Unlike the conventional fully supervised scenario, one-shot learning is particularly relevant for sensing applications, where labeled samples are expensive and difficult to acquire [15, 57]. In this setup, we select only *one labeled* sample per class from the test portion for fine-tuning on the downstream task, while the remaining test samples are used to compute the accuracy metrics reported in our results tables. This challenging setting underscores Babel's effectiveness as a pre-trained network, demonstrating its ability to generalize with minimal supervision in resource-constrained sensing environments.



a On in-domain datasets **b On out-of-domain datasets**

Figure 7: One-shot HAR classification accuracy of each modality achieved by Babel on the in-domain and out-of-domain datasets compared to various singular-modal baselines: LIMU-BERT(IMU), SenseFi(Wi-Fi) [62], MARS(mmWave) [3], PointTransformer(LiDAR), ResNet3D(Video), ST-GCN(Skeleton). For out-of-domain datasets, Babel achieves 4.5×, 7.6×, 5×, and 10.4× improvements over random guessing (IMU: 16.7%, Wi-Fi: 4.5%, mmWave: 8.3%, LiDAR: 5.0%).

We compare Babel with a broad range of baselines, including SOTA singular-modal sensing baselines, LIMU-BERT [58] for IMU, SenseFi [63] for Wi-Fi, MARS [3] for mmWave, MeteorNet [34] together with PointTransformer [72] for LiDAR. Additionally, we include the multi-modal sensing baseline Cosmo [37]. Finally, we compare Babel with emerging MLLMs that hold potential for interpreting sensing signals, including OneLLM [20] and M4 [66].

8.1.1 Performance on singular-modal sensing. Owing to the pre-trained alignment across multiple sensing modalities, Babel exhibits superior performance even when each individual aligned modality is applied to downstream tasks. In our evaluation, we apply one-shot training and testing exclusively to each modality's data—even if the dataset itself contains multiple modalities. For instance, although UTD-MHAD includes IMU, video, and skeleton data for the same set of actions, we use only the IMU samples for one-shot training and testing on the IMU, only the video samples for the video evaluation, and so forth. Fig. 7a presents the evaluation results of singular-modal sensing across four in-domain datasets. As illustrated, compared to SOTA singular-modal methods, Babel delivers an average accuracy improvement of approximately 12% across six aligned modalities in various datasets. Notably Babel achieves significant gains in weaker sensing modalities. For instance, classification accuracy for 27 human activities in the IMU modality increases from 20.19% to 31.77%, compared to LIMU-BERT on UTD-MHAD. The Wi-Fi modality obtains an approximate 10.74% enhancement. The mmWave modality shows a substantial increase from 30.32% to 50.30%, and the LiDAR modality achieves an accuracy of 43.91%, up from 28.43%. Such gains are achieved by aligning each modality into a unified representation space, facilitating

Table 2: The classification accuracy of sensing fusion for Babel and the multi-modal sensing baselines evaluated on the out-of-domain mRI[1] dataset.

	Multi-modal Baselines	Babel
Vision+IMU	89.6%	91.7%(+2.3%)
Vision+mmWave	58.4%	64.6%(+10.7%)
IMU+mmWave	75.0%	86.5%(+13.5%)
Vision+IMU+mmWave	85.4%	92.8%(+8.6%)

Table 3: Comparison with Cosmo [37] under various configurations for IMU-Skeleton fusion sensing: Babel (B) represents the Babel network that only aligns two modalities, while Cosmo(G) denotes the network customized its specific structure with simple MLPs.

	Babel(B)	Cosmo	Cosmo(G)	Babel
Aligned Modalities	2	2	2	6
Downstream Task Designs	MLP	Specific network and training strategy	MLP	MLP
IMU-Skeleton Fusion Acc.	61.46%	56.3%	41%	63.02%

mutual learning from strong modalities, *i.e.,* video. Though sensing modalities benefit significantly, gains are limited for the video modality, increasing by around 2%.

The performance of Babel on four full out-of-domain datasets is detailed in Fig. 7b. Owing to the effectiveness of multi-modality alignment, Babel consistently outperforms the SOTA methods for each individual modality. Notably, Babel demonstrates significant improvements for the mmWave and LiDAR modalities, achieving gains of 14.6% and 13.1%, respectively. In the Wi-Fi modality, Babel outperforms SenseFi by 5.3%. For the IMU modality, Babel attains an accuracy of 74.9%. It is important to note that none of the datasets evaluated here were included in the pre-training dataset, highlighting the generality of Babel.

8.1.2 Performance on multi-modal sensing. The unified representation space in Babel allows for effective fusion. When IMU and Video modalities are fused, Babel achieves a 33.17% accuracy on UTD-MHAD [9], outperforming both the individual IMU and video modalities. Likewise, a 58.97% accuracy is achieved on XRF55 [51] when Wi-Fi and mmWave modalities are merged. Note that even modality combinations (like IMU&Video fusion) not included in pre-training datasets are evaluated and obtain the superior performance, highlighting Babel's flexibility and offering developers numerous opportunities to choose any one or combined modalities for their tasks.

We also evaluate the multi-modal sensing fusion of Babel on an out-of-domain dataset, mRI[1], which is not included in the pre-training datasets. Various combinations of modalities are evaluated, and Babel's performance is compared with multi-modal sensing baselines. These baselines are implemented following the designs of existing works [2, 3]. As shown in Table 2, Babel achieves up to a 13.5% improvement in accuracy compared to the modality-specific baselines.

Table 4: HAR classification accuracy with Babel and typical MLLMs on UTD-MHAD datasets.

MLLMs	IMU UTD-MHAD [9]	Video UTD-MHAD	Wi-Fi OPERANet [4]	mmWave XRF55 [51]	Skeleton UTD-MHAD	LiDAR MM-Fi [64]
OneLLM [20] with **Meta-Transformer** [70]	6.5%	6.51%	–	–	–	–
M4 [66] with **ImageBind** [17]	5.77%	7.44%	–	–	–	–
Ours	**31.77%**	**21.35%**	**33.89%**	**50.30%**	**61.06%**	**43.91%**

8.1.3 Comparison with Cosmo [37]. Cosmo is the SOTA sensing fusion framework, but unlike Babel, it requires all modalities to coexist within one dataset, limiting its expandability to datasets with complete paired data. Thus to compared with Cosmo, we utilize the same paired IMU-skeleton from the UTD-MHAD [9] that it excels. An equal amount of data is employed to train both Cosmo and a bi-modality version of Babel. We train Cosmo and Babel for 10 times with different random seeds. In the fusion of IMU and skeleton modalities, Cosmo achieves an averaged classification accuracy of 56.3%, while Babel attains 61.46% on UTD-MHAD, as shown in Table 3.

What's more, Cosmo's performance can also be attributed to the integration of an additional network structure and its corresponding training procedure (for each downstream task), referred to as iterative fusion learning. Conversely, we aim to highlight Babel's efficacy as a pre-trained network with a simple downstream task design. When applying the same downstream network (*i.e.,* MLP), Babel could achieve around 20% accuracy improvement compared to Cosmo. Furthermore, as a expandable solution, Babel allows aligning more modalities without retraining pre-existing ones. This enhances Babel's performance when introducing modalities. As shown, Babel aligning six modalities during the pre-training phase could achieve 63.02% on the IMU-Skeleton fusion sensing task.

8.1.4 Comparison with MLLMs. There has been a significant development in MLLMs [20, 70]. These models are capable of understanding multi-modal inputs, including sensing modalities like IMU potentially. For comparison, we select typical MLLMs *e.g.,* OneLLM [20] and M4 [66], and evaluate their performance on the UTD-MHAD [9] with HAR tasks. OneLLM uses Meta-Transformer [70] and M4 uses ImageBind [17] to interpret sensing signals, respectively. The results are summarized in Table 4. Firstly, current MLLMs can only support a limited number of sensing modalities, *i.e.,* IMU. Secondly, they only achieve a classification accuracy of around 5%-6% according to our evaluation. In stark contrast, Babel significantly outperforms these with a classification accuracy of 31.77% on IMU while supporting other five sensing modalities.

The rationale that these MLLMs seems supporting sensing modalities, but struggle to comprehend IMU data and manage HAR tasks, is their training only on the Ego4D dataset [19]. Without sufficient training, these models are restricted to trained data, limiting their cross-domain capabilities. Furthermore, these MLLMs are unable to be trained on other sensing datasets due to data scarcity and absence of techniques like pre-trained modality tower and the expandable architecture, which are introduced in Babel. MLLMs could be improved by incorporating Babel as the sensing modality encoders, which would be discussed in §8.2

Table 5: Performance of each modality when applying different heuristics to determine growth orders.

Heuristics	IMU UTD-MHAD [9]	Skeleton UTD-MHAD	Video UTD-MHAD	Wi-Fi OPERANet [4]	mmWave XRF55 [51]	LiDAR MM-Fi [64]
Random	31.77%	61.06%	21.35%	33.89%	50.30%	43.91%
Robustness	29.33%	60.58%	20.83%	35.31%	52.16%	44.65%
Diversity	27.60%	56.25%	21.35%	35.79%	47.93%	44.21%
Amount	28.13%	59.90%	20.83%	33.85%	46.85%	47.70%

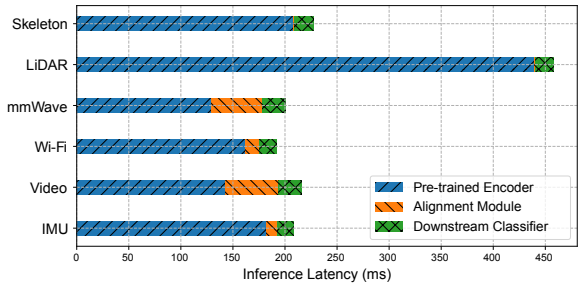


Figure 8: Per-sample inference latency breakdown for each modality in Babel evaluated on NVidia A100 GPU.

8.1.5 Growth Orders. Thanks to the prototype network architecture and the adaptive training strategy proposed in Babel, the order of modality growth does not significantly influence the end-to-end performance once the training is sufficient. To evaluate this, we devise four network growth sequences according to different heuristics: *(i)* random order: the modalities are aligned in the sequence of IMU, skeleton, video, Wi-Fi, mmWave, and LiDAR. *(ii)* alignment from the most robust to weakest modality (skeleton, video, LiDAR, IMU, Wi-Fi, mmWave); *(iii)* alignment based on data diversity, taking into account the number of actions, subjects, and scenes of the used datasets; The sequence follows skeleton, mmWave, LiDAR, IMU, and Wi-Fi. *(iv)* alignment based on the data amount of used datasets, organized from largest to smallest, proceeding from skeleton, mmWave, Wi-Fi, LiDAR, and IMU. Using these different growth orders, we train Babel and evaluate the end-to-end classification accuracy on the downstream tasks. As shown in Table 5, growth order doesn't significantly affect the performance. For instance, with different growth orders, the performance on IMU and Wi-Fi modality varies less than 3% and 2%, respectively. This highlights Babel's robustness.

8.1.6 Ablation. The techniques proposed in Babel, including the pre-trained encoders, expandable network architecture and adaptive training strategy, are all essential for constructing Babel. Particularly, Without pre-trained modality tower, training wouldn't converge due to limited samples. On UTD-MHAD [9], without prototype network, the previously aligned modality would drop about averagely 44.7% relatively after introducing a new modality. Without adaptive training, the overall performance would decrease by up to 7.2%.

8.1.7 System Overhead. In Babel, we demonstrates efficiency by introducing minimal additional system overhead. Fig. 8 illustrates the breakdown of inference latency for each modality in Babel, evaluated on a NVidia A100 GPU for an activity sample. As

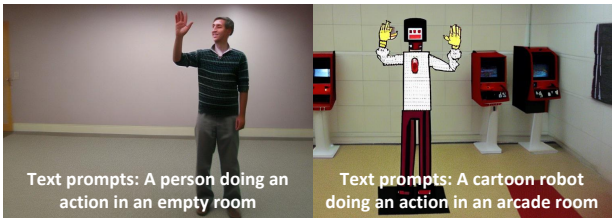


Figure 9: Images generated through cross-modality retrieval. The action information (waving hands) is input via IMU, the environment information is input through text prompts.

depicted, the overhead introduced by Babel, specifically the alignment module, is very limited compared to the pre-trained encoders. For example, in the IMU modality, the alignment module requires approximately 10.1 ms, while the pre-trained encoder, LIMU-BERT, takes 182.4 ms to process a fixed window of IMU signals. On average, the alignment module incurs only about an 8% increase in inference overhead based on our evaluation. The overhead of the alignment module varies across modalities due to the use of different MLP layer configurations to accord with each modality's encoder. Aside from the alignment module, the prototype network is shared among modalities, with its overhead being negligible, less than 1 ms. When applying multiple modalities to downstream tasks, each modality tower can be parallelized to hide multi-modal sensing latency.

The pre-trained weights of Babel occupy approximately 1.1GB on disk, encompassing pre-trained encoders for all modalities, concept alignment modules, and the prototype network. Depending on the selected modalities, Babel requires 1.4-9.92GB of memory using FP32 precision.

8.2 Case Study

8.2.1 Cross-modality retrieval. The alignment of diverse sensing modalities in Babel potentially opens up the possibilities for cross-modality retrieval applications. This involves obtaining the representations of one modality using signals from other modalities as inputs. Such applications could be promising. For instance, using wireless sensing signals as input to retrieve visual representations could be considered an example of *sensing imaging*.

To showcase, we construct a prototype designed to retrieve visual representations and generate images using non-visual sensors, such as IMU. Specifically, we align Babel with unCLIP [41], an image-to-image diffusion model. unCLIP employs an image encoder to obtain the embeddings of the input image and then uses these embeddings to guide the diffusion process, thereby generating images that bear stylistic similarities to the input image. We incorporate unCLIP's image encoder into our Babel network, enabling the sensing modalities to be interpreted by the diffusion module in unCLIP. We use L1 loss to align Babel and unCLIP.

Fig. 9 demonstrates the images generated using IMU as inputs, representing the sensor readings of a person gesturing with hands. Leveraging unCLP, the actions captured by IMU are visually represented. The environmental information and other visual styles are provided through text prompts. We believe this area of research

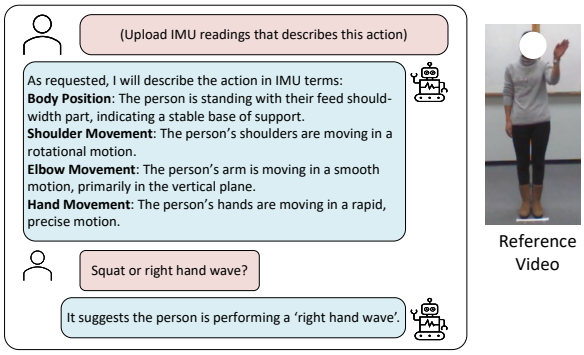


Figure 10: With Babel, Video-LLaMA accepts IMU readings as inputs and conduct a preliminary analysis of the actions represented by these IMU readings.

opens up interesting possibilities, offering a pathway to visualizing the physical world through non-visual sensors.

8.2.2 Bridge with LLMs. The alignment of diverse sensing modalities into a unified representation presents an advantageous prospect for integration with LLMs. To demonstrate, we integrate Babel with Video-LLaMA [68], which is a multi-modal LLM with the ability to understand both visual and audio contents.

We establish the alignment between the video modality in Babel and that in Video-LLaMA. Specifically, we judiciously select the video encoder from Video-LLaMA and construct a modality tower for integration into Babel. We employ the L1 loss in this scenario, ensuring the video encoder of Video-LLaMA remains frozen while all modalities in Babel align towards Video-LLaMA. This strategy aims to generate embeddings of sensing modalities that could potentially be interpreted by Video-LLaMA.

Fig. 10 provides an impressive illustration where we input an IMU sequence depicting a woman waving her hands. These IMU readings are processed by Babel and subsequently fed into Video-LLaMA. Remarkably, without any specific training on LLMs, it successfully deciphers the action captured by the IMU data and, when promoted, differentiates between diverse actions, such as squatting or waving hands. This exemplifies the potential of bridging sensing and LLMs via the modality alignment introduced by Babel. Our future research will concentrate on improving Babel, aiming to bolster the model's capability to provide deeper insights and more accurate interpretations of physical world based on a broader spectrum of sensing modalities, and bringing such the capabilities to LLMs.

9 DISCUSSION AND FUTURE WORK

While Babel demonstrates promising results in aligning multiple sensing modalities for HAR tasks, several directions remain for our future exploration. Firstly, the framework could be extended to other sensing applications including localization, gesture detection [46], and autonomous navigation. Secondly, Our preliminary exploration of integrating Babel with LLMs demonstrates the potential for enhanced sensing comprehension. The current approach of aligning sensing embeddings with video encoders in MLLMs. Future research should investigate direct integration methods that can map sensing features into LLM's native representation spaces without requiring intermediate alignment steps. Our case studies in sensing imaging and LLM integration could be further benefit from more state-of-the-art diffusion models [21, 54, 56] and LLMs [26, 29, 69].

10 CONCLUSION

We present Babel, a expandable modality alignment model designed for sensing applications. The pre-trained Babel has been proficiently aligned with six prevalent sensing modalities, IMU, skeleton, video, Wi-Fi, LiDAR, and mmWave. Babel demonstrated the superior performance for HAR tasks across various datasets compared to an array of baselines. As Babel is a scalable network, we call for the community to further enhance and align additional helpful modalities into Babel.

ACKNOWLEDGMENTS

Authors are supported in part by the following: US National Science Foundation awards — CNS-2112562, CNS-2107060, CNS-2213688, CNS-2312716, and the US Department of Commerce award 70NANB21H043.

REFERENCES

[1] Sizhe An, Yin Li, and Umit Ogras. 2022. mRI: Multi-modal 3D Human Pose Estimation Dataset using mmWave, RGB-D, and Inertial Sensors. arXiv:2210.08394 [cs.CV] https://arxiv.org/abs/2210.08394
[2] Sizhe An, Yin Li, and Umit Ogras. 2022. mri: Multi-modal 3d human pose estimation dataset using mmwave, rgb-d, and inertial sensors. *Advances in Neural Information Processing Systems* 35 (2022), 27414–27426.
[3] Sizhe An and Umit Y Ogras. 2021. Mars: mmwave-based assistive rehabilitation system for smart healthcare. *ACM Transactions on Embedded Computing Systems (TECS)* 20, 5s (2021), 1–22.
[4] Mohammud J. Bocus, Wenda Li, Shelly Vishwakarma, Roget Kou, Chong Tang, Karl Woodbridge, Ian Craddock, Ryan McConville, Raul Santos-Rodriguez, Kevin Chetty, and Robert Piechocki. 2021. OPERAnet: A Multimodal Activity Recognition Dataset Acquired from Radio Frequency and Vision-based Sensors. arXiv:2110.04239 [eess.SP]
[5] Leo Breiman. 1996. Bagging predictors. *Machine learning* 24 (1996), 123–140.
[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7291–7299.
[7] Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. 2004. Ensemble selection from libraries of models. In *Proceedings of the twenty-first international conference on Machine learning*. 18.
[8] Anjun Chen, Xiangyu Wang, Shaohao Zhu, Yanxu Li, Jiming Chen, and Qi Ye. 2023. mmBody Benchmark: 3D Body Reconstruction Dataset and Analysis for Millimeter Wave Radar. arXiv:2209.05070 [cs.CV]
[9] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. 2015. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 IEEE International conference on image processing (ICIP)*. IEEE, 168–172.
[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.
[11] Seungeun Chung, Jiyoun Lim, Kyoung Ju Noh, Gague Kim, and Hyuntae Jeong. 2019. Sensor data acquisition and multimodal sensor fusion for human activity recognition using deep learning. *Sensors* 19, 7 (2019), 1716.
[12] 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 (2022), 108:1–108:28.
[13] Yiwen Dong, Jingxiao Liu, and Hae Young Noh. 2022. GaitVibe+: Enhancing Structural Vibration-Based Footstep Localization Using Temporary Cameras for in-Home Gait Analysis. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems (SenSys '22)*. ACM.
[14] Li Fei-Fei, Robert Fergus, and Pietro Perona. 2006. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence* 28, 4 (2006), 594–611.
[15] Siwei Feng and Marco F Duarte. 2019. Few-shot learning-based human activity recognition. *Expert Systems with Applications* 138 (2019), 112782.

[16] Robert M French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences* 3, 4 (1999), 128–135.

[17] 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]

[18] Lewis Girod and Deborah Estrin. 2001. Robust range estimation using acoustic and multimodal sensing. In *Proceedings 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems. Expanding the Societal Role of Robotics in the the Next Millennium (Cat. No. 01CH37180)*, Vol. 3. IEEE, 1312–1320.

[19] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18995–19012.

[20] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. 2023. OneLLM: One Framework to Align All Modalities with Language. arXiv:2312.03700 [cs.CV]

[21] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. 2024. Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation. arXiv:2311.17117 [cs.CV] https://arxiv.org/abs/2311.17117

[22] Isibor Kennedy Ihianle, Augustine O Nwajana, Solomon Henry Ebenuwa, Richard I Otuka, Kayode Owa, and Mobolaji O Orisatoki. 2020. A deep learning approach for human activities recognition from multimodal sensing devices. *IEEE Access* 8 (2020), 179028–179038.

[23] Yash Jain, Chi Ian Tang, Chulhong Min, Fahim Kawsar, and Akhil Mathur. 2022. ColloSSL: Collaborative Self-Supervised Learning for Human Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 1 (2022), 17:1–17:28.

[24] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The Kinetics Human Action Video Dataset. arXiv:1705.06950 [cs.CV]

[25] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, Vol. 2. Lille, 1–30.

[26] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. LLaVA-OneVision: Easy Visual Task Transfer. arXiv:2408.03326 [cs.CV] https://arxiv.org/abs/2408.03326

[27] Jiaxin Li, Danfeng Hong, Lianru Gao, Jing Yao, Ke Zheng, Bing Zhang, and Jocelyn Chanussot. 2022. Deep learning in multimodal remote sensing data fusion: A comprehensive review. *International Journal of Applied Earth Observation and Geoinformation* 112 (2022), 102926.

[28] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. 2010. Action recognition based on a bag of 3d points. In *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*. IEEE, 9–14.

[29] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122* (2023).

[30] Hansi Liu, Abrar Alali, Mohamed Ibrahim, Hongyu Li, Marco Gruteser, Shubham Jain, Kristin Dana, Ashwin Ashok, Bin Cheng, and Hongsheng Lu. 2021. Lost and Found! associating target persons in camera surveillance footage with smartphone identifiers. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '21)*. Association for Computing Machinery, New York, NY, USA, 499–500.

[31] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning. arXiv:2205.05638 [cs.LG]

[32] Shengzhong Liu, Tomoyoshi Kimura, Dongxin Liu, Ruijie Wang, Jinyang Li, Suhas Diggavi, Mani Srivastava, and Tarek Abdelzaher. 2024. FOCAL: Contrastive learning for multimodal time-series sensing signals in factorized orthogonal latent space. *Advances in Neural Information Processing Systems* 36 (2024).

[33] Xiulong Liu, Dongdong Liu, Jiuwu Zhang, Tao Gu, and Keqiu Li. 2021. RFID and camera fusion for recognition of human-object interactions. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 296–308.

[34] Xingyu Liu, Mengyuan Yan, and Jeannette Bohg. 2019. Meteornet: Deep learning on dynamic 3d point cloud sequences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9246–9255.

[35] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101 [cs.LG]

[36] Arun Mallya and Svetlana Lazebnik. 2018. PackNet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7765–7773.

[37] Xiaomin Ouyang, Xian Shuai, Jiayu Zhou, Ivy Wang Shi, Zhiyuan Xie, Guoliang Xing, and Jianwei Huang. 2022. Cosmo: contrastive fusion learning with small data for multimodal human activity recognition. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*. 324–337.

[38] Quang Pham, Chenghao Liu, and Steven Hoi. 2021. DualNet: Continual learning, fast and slow. *Advances in Neural Information Processing Systems* 34 (2021), 16131–16144.

[39] 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]

[40] Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D Lane, Cecilia Mascolo, Mahesh K Marina, and Fahim Kawsar. 2018. Multimodal deep learning for activity and context recognition. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 1, 4 (2018), 1–27.

[41] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125 [cs.CV]

[42] Jorge-L Reyes-Ortiz, Luca Oneto, Albert Samà, Xavier Parra, and Davide Anguita. 2016. Transition-aware human activity recognition using smartphones. *Neurocomputing* 171 (2016), 754–767.

[43] Elan Rosenfeld, Preetum Nakkiran, Hadi Pouransari, Oncel Tuzel, and Fartash Faghri. 2022. APE: Aligning Pretrained Encoders to Quickly Learn Aligned Multimodal Representations. arXiv:2210.03927 [cs.LG]

[44] Batool Salehi, Guillem Reus-Muns, Debashri Roy, Zifeng Wang, Tong Jian, Jennifer Dy, Stratis Ioannidis, and Kaushik Chowdhury. 2022. Deep learning on multimodal sensor data at the wireless edge for vehicular network. *IEEE Transactions on Vehicular Technology* 71, 7 (2022), 7639–7655.

[45] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. 2016. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. arXiv:1604.02808 [cs.CV]

[46] Yuliang Sun, Tai Fei, Xibo Li, Alexander Warnecke, Ernst Warsitz, and Nils Pohl. 2020. Real-time radar-based gesture detection and recognition built in an edge-computing platform. *IEEE Sensors Journal* 20, 18 (2020), 10706–10716.

[47] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. 2023. Any-to-Any Generation via Composable Diffusion. arXiv:2305.11846 [cs.CV] https://arxiv.org/abs/2305.11846

[48] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2019. Contrastive Multiview Coding. *CoRR* abs/1906.05849 (2019). arXiv:1906.05849 http://arxiv.org/abs/1906.05849

[49] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A Closer Look at Spatiotemporal Convolutions for Action Recognition. arXiv:1711.11248 [cs.CV]

[50] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *CoRR* abs/1807.03748 (2018). arXiv:1807.03748 http://arxiv.org/abs/1807.03748

[51] Fei Wang, Yizhe Lv, Mengdie Zhu, Han Ding, and Jinsong Han. 2024. XRF55: A Radio Frequency Dataset for Human Indoor Action Analysis. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 1, Article 21 (mar 2024), 34 pages.

[52] Ziwei Wang, Jiajun Liu, Reza Arablouei, Greg Bishop-Hurley, Melissa Matthews, and Paulo Borges. 2022. Multi-modal sensing for behaviour recognition. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*. 900–902.

[53] Yuxuan Weng, Guoquan Wu, Tianyue Zheng, Yanbing Yang, and Jun Luo. 2024. Large Model for Small Data: Foundation Model for Cross-Modal RF Human Activity Recognition. arXiv:2410.19766 [cs.CV] https://arxiv.org/abs/2410.19766

[54] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2023. Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation. arXiv:2212.11565 [cs.CV] https://arxiv.org/abs/2212.11565

[55] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 2015. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1912–1920.

[56] Zhiqiang Xia, Zhaokang Chen, Bin Wu, Chao Li, Kwok-Wai Hung, Chao Zhan, Yingjie He, and Wenjiang Zhou. 2024. MuseV: Infinite-length and High Fidelity Virtual Human Video Generation with Visual Conditioned Parallel Denoising. *arxiv* (2024).

[57] Rui Xiao, Jianwei Liu, Jinsong Han, and Kui Ren. 2021. Onefi: One-shot recognition for unseen gesture via cots wifi. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 206–219.

[58] Huatao Xu, Pengfei Zhou, Rui Tan, Mo Li, and Guobin Shen. 2021. Limu-bert: Unleashing the potential of unlabeled data for imu sensing applications. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 220–233.

[59] Huatao Xu, Pengfei Zhou, Rui Tan, Mo Li, and Guobin Shen. 2022. LIMU-BERT: Unleashing the Potential of Unlabeled Data for IMU Sensing Applications. *GetMobile: Mobile Computing and Communications* 26, 3 (2022), 39–42.

[60] Lilin Xu, Chaojie Gu, Rui Tan, Shibo He, and Jiming Chen. 2023. MESEN: Exploit Multimodal Data to Design Unimodal Human Activity Recognition with Few Labels. (2023).

[61] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. arXiv:1801.07455 [cs.CV]

[62] Jianfei Yang, Xinyan Chen, Dazhuo Wang, Han Zou, Chris Xiaoxuan Lu, Sumei Sun, and Lihua Xie. 2023. SenseFi: A Library and Benchmark on Deep-Learning-Empowered WiFi Human Sensing. arXiv:2207.07859 [cs.LG]

[63] Jianfei Yang, Xinyan Chen, Dazhuo Wang, Han Zou, Chris Xiaoxuan Lu, Sumei Sun, and Lihua Xie. 2023. SenseFi: A Library and Benchmark on Deep-Learning-Empowered WiFi Human Sensing. arXiv:2207.07859 [cs.LG] https://arxiv.org/abs/2207.07859

[64] Jianfei Yang, He Huang, Yunjiao Zhou, Xinyan Chen, Yuecong Xu, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie. 2023. MM-Fi: Multi-Modal Non-Intrusive 4D Human Dataset for Versatile Wireless Sensing. arXiv:2305.10345 [eess.SP]

[65] Siamak Yousefi, Hirokazu Narui, Sankalp Dayal, Stefano Ermon, and Shahrokh Valaee. 2017. A survey on behavior recognition using WiFi channel state information. *IEEE Communications Magazine* 55, 10 (2017), 98–104.

[66] Jinliang Yuan, Chen Yang, Dongqi Cai, Shihe Wang, Xin Yuan, Zeling Zhang, Xiang Li, Dingge Zhang, Hanzi Mei, Xianqing Jia, Shangguang Wang, and Mengwei Xu. 2024. Mobile Foundation Model as Firmware. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking* (Washington D.C., DC, USA) *(ACM MobiCom '24)*. Association for Computing Machinery, New York, NY, USA, 279–295. https://doi.org/10.1145/3636534.3649361

[67] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. LiT: Zero-Shot Transfer with Locked-image text Tuning. arXiv:2111.07991 [cs.CV]

[68] 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]

[69] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. 2024. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320* (2024).

[70] Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. 2023. Meta-Transformer: A Unified Framework for Multimodal Learning. arXiv:2307.10802 [cs.CV]

[71] Yi Zhang, Yue Zheng, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. 2022. Widar3.0: Zero-Effort Cross-Domain Gesture Recognition With Wi-Fi. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 11 (2022), 8671–8688. https://doi.org/10.1109/TPAMI.2021.3105387

[72] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip H. S. Torr, and Vladlen Koltun. 2020. Point Transformer. *CoRR* abs/2012.09164 (2020). arXiv:2012.09164 https://arxiv.org/abs/2012.09164

[73] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. 2018. Through-Wall Human Pose Estimation Using Radio Signals. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR '18)*.

[74] Pengfei Zhou, Mo Li, and Guobin Shen. 2014. Use it free: instantly knowing your phone attitude. In *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking (MobiCom '14)*.

[75] Yunjiao Zhou, Jianfei Yang, Han Zou, and Lihua Xie. 2023. TENT: Connect Language Models with IoT Sensors for Zero-Shot Activity Recognition. arXiv:2311.08245 [cs.CV] https://arxiv.org/abs/2311.08245

[76] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Li Yuan. 2024. LanguageBind: Extending Video-Language Pretraining to N-modality by Language-based Semantic Alignment. arXiv:2310.01852 [cs.CV] https://arxiv.org/abs/2310.01852