# Chained Cross-Modal Learning for Modality-Agnostic Embeddings

**Naoki Tamura**[1,2,*] **Taiki Sekii**[1,†]
[1]CyberAgent
[2]*Graduate School of Engineering*, Nagoya University
taiki.sekii@gmail.com

## Abstract

This study proposes a cross-modal learning framework that learns modality-agnostic embeddings for six modalities: language, temporal 2D/3D skeletons, IMU, GPS, and Wi-Fi. Existing approaches such as ImageBind align each target modality with images. Consequently, they are restricted to modalities for which paired image data can be prepared a priori, which makes them unsuitable for time-series tasks, such as action recognition. To overcome this limitation, we propose *chained modal alignment*, which aligns the six modalities sequentially via temporal 3D skeletons, thereby enabling emergent zero-shot action recognition based on a shared embedding space. Because the proposed method relaxes the requirement of direct pairing with images, it supports a wider range of modality combinations and thus expands practical applications. In addition, we introduce *chained modal transfer learning*, which transfers the knowledge learned between language and 3D skeletons to other modalities with data that cannot be directly paired, thereby improving generalization. Experiments on three public datasets spanning all six modalities quantitatively validate the effectiveness of the proposed method for emergent zero-shot action recognition.

## 1 Introduction

Statistical language modeling has advanced dramatically, culminating in large language models (LLMs) (Touvron et al., 2023; OpenAI et al., 2024) that underpin numerous applications (OpenAI; GitHub, Inc.). A key driver of this progress is the evolution of deep neural architectures, as exemplified by Transformers (Vaswani et al., 2017; Touvron et al., 2023; OpenAI et al., 2024) that handle unstructured data of varying granularity, such as tokens, words, and sentences. The versatility of Transformers has made them de facto architectures across modalities beyond language (Dosovitskiy et al., 2021; Huang et al., 2018). Consequently, the relationships between text and data from other modalities can be modeled, with the rapid advancements leading to multimodal LLMs capable of visual question answering and video generation (Runway AI, Inc.; OpenAI; Gemini Team, 2025).

Among all modalities, vision and audio have benefited the most from LLM advances, largely because their data can be readily paired with text (*e.g.*, captions and transcripts) harvested from web services. With abundant paired data, each modality can be embedded into a linguistic feature space interpretable by LLMs through a process we call *modal alignment*. Contrastive learning is typically employed to align the feature representations of each modality with the corresponding textual embeddings. Subsequently, modality-specific encoders are fused with an LLM to build a multimodal LLM (Chen et al., 2024; Bai et al., 2025).

### 1.1 Limitations of Prior Work

Focusing on human action understanding, commonly applied in domains such as sports and education (Wu et al., 2023), we study the zero-shot action recognition task, in which the system classifies actions in response to a user's text query without having seen labeled examples of these actions. Recent studies (Tevet et al., 2022; Moon et al., 2023; Vivanco Cepeda et al., 2023) leverage diverse modalities, including inertial measurement unit (IMU), global positioning system (GPS), and Wi-Fi.[1] However, unlike vision or audio, these modalities rarely come with large-scale text pairs obtainable from the internet. As with multimodal LLMs, enabling zero-shot recognition requires collecting numerous paired datasets per modality or sensor, which is a prohibitively expensive endeavor

[*] The work was conducted at CyberAgent, Inc.
[†] Corresponding author.
[1] This refers to Wi-Fi channel state information (CSI).



*Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026*, pages 3710–3723
November 6-10, 2026 ©2026 Association for Computational Linguistics

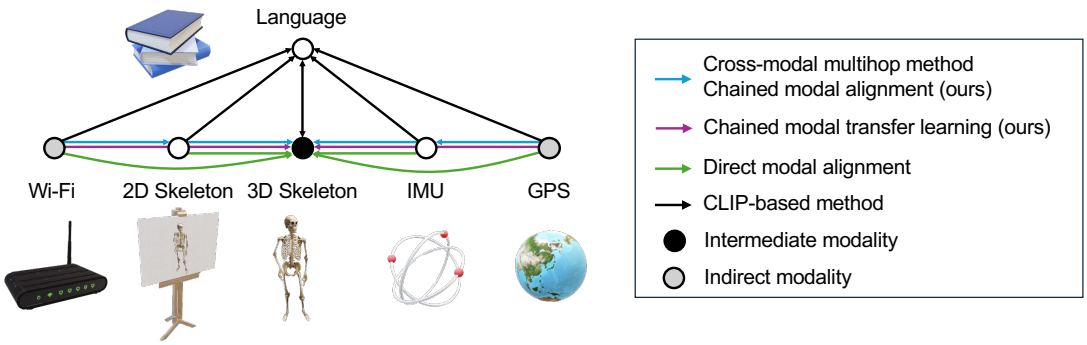



Figure 1: Relationship between the proposed framework and the baselines.

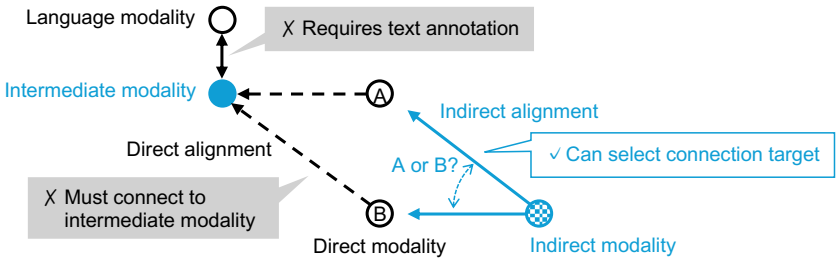



Figure 2: Direct modal alignment in prior work (Girdhar et al., 2023; Moon et al., 2023) compared with the proposed indirect modal alignment. Our method allows the DNN of an *indirect modality* to share the embedding space with an intermediate modality without training on paired data between them.

in both sensing and textual annotation. Accordingly, this remains an open problem across many fields in which LLMs are expected to excel (Xu et al., 2025).

To mitigate data-collection costs, unsupervised modal alignment methods (Girdhar et al., 2023; Moon et al., 2023) have recently been proposed. These align multiple modalities to an *intermediate modality* (typically images) aligned with the language. Given the paired data between the intermediate modality and each *target modality*,[2] they embed all modalities into a shared space without any textual annotations for the target modalities. Our goal is to leverage these benefits for modalities used in action recognition and address the two challenges of existing approaches:

**Challenge 1: Extending to time-series modality pairs.** Using images as the intermediate modality confines the learned representation to what can be inferred from static frames.[3] In contrast, action recognition relies on sequential data with privacy concerns, low-light conditions, and high acquisition costs, further motivating the exploration of non-vision modalities.

**Challenge 2: Flexible selection of modality pairs.** Existing methods require every target modality to be directly aligned with an intermediate modality. In this paper, we call this scheme *direct modal alignment* (see Figs. 1 and 2). Modalities that cannot be paired with images are thus excluded, limiting their applicability. Therefore, a more flexible strategy for selecting modality pairs is desirable.

### 1.2 Overview and Contributions

To jointly resolve Challenges 1 and 2, we present a new cross-modal learning framework that targets six modalities relevant to action recognition: language, temporal 2D/3D skeletons, IMU, GPS, and Wi-Fi. For Challenge 1, we designate 3D skeletons as the intermediate modality and embed all six modalities into a shared embedding space capable of representing time-series actions. As illustrated in Fig. 1, we define modality pairs in a chain, *e.g.*, 3D skeleton–IMU, IMU–GPS or 3D–2D skeleton, 2D skeleton–Wi-Fi, according to how easily paired samples can be collected and how informative each modality is regarding human motion (*cf.* Sec. 3.1). It introduces *indirect modalities* (*e.g.*, GPS, Wi-

[2]Here, the term *target modalities* refers to all modalities other than intermediate and language modalities.

[3]Although prior work (Girdhar et al., 2023; Moon et al., 2023) mentions videos as a potential intermediate modality, it actually samples only a few sparse frames.

Fi) that are aligned indirectly with the intermediate modality via another modality. To the best of our knowledge, this is the first study to explore indirect modal alignment for time-series action recognition without target–intermediate paired data. Hence, when adding a new indirect modality, one can choose whichever already-aligned modality is easiest to pair with, which dramatically expands the possible alignments (*cf.* Fig. 2) and greatly improves the practicality of real-world applications. For instance, IMU and GPS data can be recorded simultaneously on a single device, eliminating the need to associate indirect modalities with images and reducing data-collection costs.

We further propose *chained modal transfer learning*, which transfers knowledge sequentially from an intermediate modality to an indirect modality along a path, either language $\rightarrow$ GPS or language $\rightarrow$ Wi-Fi, by jointly optimizing all the alignments on that path. Unlike conventional cross-modal transfer learning, this approach simultaneously achieves indirect alignment and eliminates the learning errors that accumulate as the modality path from language to indirect modalities grows longer.

Using representative public datasets for each modality, we quantitatively evaluated emergent zero-shot action recognition (Girdhar et al., 2023). Without any data directly paired with the intermediate modality, our method attained accuracy on indirect modalities comparable to that of direct-alignment baselines, suggesting that error accumulation introduced by chained alignment is negligible in practice under our setting.

Our contributions are fourfold, demonstrating modal alignment across four or more modalities in that (1) 3D skeletons serve as effective intermediate modalities for time-series tasks such as action recognition. (2) Chaining target modalities indirectly to an intermediate modality enables shared embedding across heterogeneous modalities. (3) Chained modal transfer learning improves the generalization of indirect-modality DNNs by transferring knowledge from the intermediate modality. (4) We conduct a comprehensive evaluation across pivot choices, chain topologies, and path lengths, validating two end-modality routes (GPS and Wi-Fi).

## 2 Related Work

### 2.1 Cross-Modal Embedding Learning

This study defines *unsupervised modal alignment* as an alignment process that requires neither textual annotations for target-modality data nor manual pairing of data across modalities. ImageBind (Girdhar et al., 2023) aligned the embedding space of each target modality (*e.g.*, audio and IMU) with that of an image encoder previously aligned with language, thereby learning a shared embedding space (see Fig. 1). IMU2CLIP (Moon et al., 2023) was proposed as an unsupervised training method that aligns language, video, and IMU data in a single embedding space.

Although many CLIP-based methods focus on contrastive learning, recent approaches leverage LLMs with high representational capacity for cross-modal learning. AnyGPT (Zhan et al., 2024) uses a pretrained LLM to model images, audio, and music within an autoregressive framework. OneLLM (Han et al., 2024) provides a framework in which an image encoder aligned with a language encoder serves as a gateway for handling data from seven modalities—including video, audio, and IMU—via a single LLM. Most of these methods use textual annotations associated with the data to embed different modalities into a shared latent space. In contrast, the proposed method aligns the indirect modality with the intermediate modality via a different target modality (other than the intermediate one), substantially increasing the number of feasible alignment combinations. A more detailed discussion of related work is provided in Appendix A.

## 3 Method

### 3.1 Overview

Figs. 1 and 3 illustrate the overall concept of the proposed framework. The framework adopts 3D skeletons as the intermediate modality and proposes two modality paths, one ending in GPS and the other ending in Wi-Fi. A key characteristic of the proposed method is that contrastive learning for each pair of modalities is arranged in a chained manner along these paths. The modality pairs considered here satisfy the following relationships: (1) motion-capture systems that can simultaneously record 3D skeletons and IMU data are now widespread; (2) a single mobile device equipped with an IMU and GPS can acquire both modalities

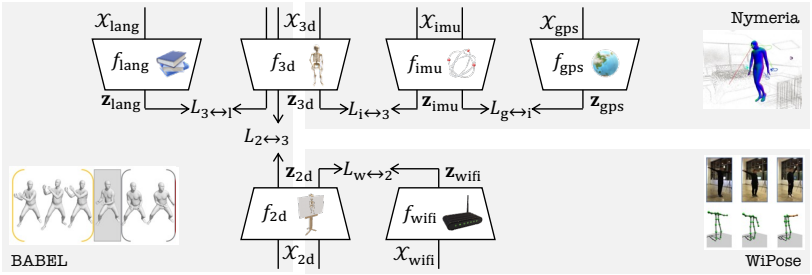



Figure 3: Overview of the proposed method.

concurrently over large areas; (3) in principle, 2D skeletons can be projected from 3D skeletons; and (4) in practical deployments, Wi-Fi and 2D skeleton data can be captured simultaneously by colocating a Wi-Fi device and a camera in the same sensing setup. The modality pairs vary in information content, and numerous studies (Zhou et al., 2023) have explored reconstructing one modality from another. The proposed method leverages these relationships to define modality pairs.

Modal alignment is first performed between the language and 3D skeleton modalities, as described in Appendix B.1. This alignment ensures that language embeddings capture the spatiotemporal features related to human actions. Each modality has an encoder $f$ parameterized by $\theta$, which embeds the input data into a shared embedding $\mathbf{z} \in \mathbb{R}^D$, where $D$ is the dimensionality of the embeddings. Thus, once the training is completed for the two paths, the classification or retrieval tasks can be solved by comparing the embeddings from any pair of modalities.

The proposed method does not require that every target modality be aligned with the intermediate modality, nor do the paired alignments share the same training data. Consequently, our method demands less manual annotation of textual data or pairing with images compared to prior studies. It also involves fewer constraints when collecting training samples. For example, when adding GPS to a modal alignment comprising language, 3D skeletons, and IMU, as the alignment counterpart, one can choose the IMU, which can provide paired data simply by carrying a single device such as a smartphone during arbitrary activities, thus obtaining a shared embedding space without multiple devices, textual annotations, or manual pairing.

## 3.2 Chained Modal Alignment

The encoders for the six modalities defined in the proposed framework are trained using two methods. In the first, each pair of modalities is aligned separately; we call this chained modal alignment. The second method, chained modal transfer learning, is described in Sec. 3.3. Specifically, language is first aligned with the 3D skeletons. The parameters of the 3D skeleton encoder are then fixed and separate contrastive learning is performed for the IMU–3D skeleton and 2D–3D skeleton pairs. Finally, the two encoders are fixed and contrastive learning is performed for the GPS–IMU and Wi-Fi–2D skeleton pairs. The appendices provide pseudocode for the training procedure (*cf.* Appendix B).

### 3.2.1 Modal Alignment for Each Modality

**3D Skeletons $\leftrightarrow$ Language.** Let $\mathcal{X}^n_{\text{3d}}$ be a time-series 3D skeleton sequence paired with a tokenized sentence $\mathcal{X}^n_{\text{lang}}$. We consider the training sample to be $\left(\mathcal{X}^n_{\text{3d}}, \mathcal{X}^n_{\text{lang}}\right)$ where $n \in \{1, \ldots, N\}$ denotes the sample index. Recent systems adopt markerless motion capture for the high-precision 3D skeleton data (Scataglini et al., 2024). $\mathcal{X}^n_{\text{3d}}$ is a time-series segment of the skeletal motion of a single individual, with joint angles represented by a 3D rotation, as in previous studies (Tevet et al., 2022). Accordingly, text tokens $\mathcal{X}^n_{\text{lang}}$ annotate the actions described by $\mathcal{X}^n_{\text{3d}}$. The pair $\left(\mathcal{X}^n_{\text{3d}}, \mathcal{X}^n_{\text{lang}}\right)$ is encoded by two DNN encoders, $f_{\text{3d}}(\cdot)$ and $f_{\text{lang}}(\cdot)$, generating embeddings $\left(\mathbf{z}^n_{\text{3d}}, \mathbf{z}^n_{\text{lang}}\right)$. Contrastive learning (Radford et al., 2021) is used to maximize the similarity for positive pairs $\left(\mathbf{z}^n_{\text{3d}}, \mathbf{z}^n_{\text{lang}}\right)$ and minimize it for negative pairs $\left(\mathbf{z}^m_{\text{3d}}, \mathbf{z}^n_{\text{lang}}\right)$ (where $m \neq n$). In what follows, we use the same contrastive learning objective for each modality pair.

**IMU $\leftrightarrow$ 3D Skeletons.** By measuring the IMU data $\mathcal{X}^n_{\text{imu}}$ along with the 3D skeleton data $\mathcal{X}^n_{\text{3d}}$, the training samples $(\mathcal{X}^n_{\text{imu}}, \mathcal{X}^n_{\text{3d}})$ for IMU–3D skeleton alignment are obtained. $\mathcal{X}^n_{\text{imu}}$ consists of 6D time-series signals (3D acceleration and 3D angular

velocities) and is transformed into an embedding $\mathbf{z}^n_{\text{imu}}$ by the encoder $f_{\text{imu}}(\cdot)$.

**GPS $\leftrightarrow$ IMU.** By measuring the GPS data $\mathcal{X}^n_{\text{gps}}$ along with the IMU data $\mathcal{X}^n_{\text{imu}}$, the training samples $\left(\mathcal{X}^n_{\text{gps}}, \mathcal{X}^n_{\text{imu}}\right)$ for the GPS–IMU alignment are obtained; $\mathcal{X}^n_{\text{gps}}$ is a 6D time-series signal representing 2D position, velocity, acceleration, heading direction, and angular velocity, and is transformed into embedding $\mathbf{z}^n_{\text{gps}}$ by the encoder $f_{\text{gps}}(\cdot)$. In this setup, $f_{\text{gps}}$ is trained only through alignment with its IMU counterpart. Consequently, although data pairs of the intermediate modality and GPS are not used for training, $f_{\text{gps}}$ can embed $\mathcal{X}^n_{\text{gps}}$ into the common embedding space shared with the intermediate modality.

**2D $\leftrightarrow$ 3D Skeletons.** By measuring the time-series 2D skeleton data $\mathcal{X}^n_{\text{2d}}$ along with the 3D skeleton data $\mathcal{X}^n_{\text{3d}}$, training samples $(\mathcal{X}^n_{\text{2d}}, \mathcal{X}^n_{\text{3d}})$ are obtained for 2D–3D skeleton alignment. Here, $\mathcal{X}^n_{\text{2d}}$ is a single-sequence 2D skeletal representation in which each joint is described by image coordinates. $\mathcal{X}^n_{\text{2d}}$ is embedded into $\mathbf{z}^n_{\text{2d}}$ using the encoder $f_{\text{2d}}(\cdot)$.

**Wi-Fi $\leftrightarrow$ 2D Skeletons.** By measuring Wi-Fi CSI data $\mathcal{X}^n_{\text{wifi}}$ along with time-series 2D skeleton data $\mathcal{X}^n_{\text{2d}}$, training samples $(\mathcal{X}^n_{\text{wifi}}, \mathcal{X}^n_{\text{2d}})$ are obtained for Wi-Fi–2D skeleton alignment. Following previous studies (Zhou et al., 2023), $\mathcal{X}^n_{\text{wifi}}$ is first converted into a time-series signal and then encoded into $\mathbf{z}^n_{\text{wifi}}$ by using the encoder $f_{\text{wifi}}(\cdot)$.

Unlike 3D skeletons, 2D skeletons can be captured both indoors and outdoors using a monocular camera without external parameter calibration. Therefore, by jointly carrying a monocular camera and a Wi-Fi transmitter/receiver in the measurement environment, one can collect training sample pairs without a motion-capture system. After completing modal alignment, an application that uses Wi-Fi data can perform emergent zero-shot action recognition using Wi-Fi devices alone, thereby avoiding personal data collection via cameras.

### 3.3 Chained Modal Transfer Learning

Chained modal transfer learning uses the encoders from chained modal alignment for initialization. The modality encoders along each of the two paths are simultaneously fine-tuned: the 3D skeleton to GPS and 3D skeleton to Wi-Fi. We let $\mathcal{L}_{\text{a}\leftrightarrow\text{b}}$ denote the contrastive loss function for modalities a and b in the contrastive framework (Radford et al., 2021). Specifically, we use $\mathcal{L}_{\text{i}\leftrightarrow 3}$ and $\mathcal{L}_{\text{g}\leftrightarrow\text{i}}$ for the IMU–3D skeleton and GPS–IMU alignments, respectively, where the subscript indicates the initial character of each modality. For the path ending in GPS, we define

$$\mathcal{L}^{\text{joint}}_{\text{g}\leftrightarrow 3} = \alpha\mathcal{L}_{\text{i}\leftrightarrow 3} + (1-\alpha)\mathcal{L}_{\text{g}\leftrightarrow\text{i}}, \tag{1}$$

and simultaneously optimize the parameters $\theta_{\text{imu}}$ and $\theta_{\text{gps}}$. $\alpha$ is the weighting coefficient of the loss terms. For the gradient of $\mathcal{L}^{\text{joint}}_{\text{g}\leftrightarrow 3}$ with respect to $\theta_{\text{imu}}$ and $\theta_{\text{gps}}$, we obtain

$$\frac{\partial\mathcal{L}^{\text{joint}}_{\text{g}\leftrightarrow 3}}{\partial\theta_{\text{imu}}} = \alpha\frac{\partial\mathcal{L}_{\text{i}\leftrightarrow 3}}{\partial\theta_{\text{imu}}} + (1-\alpha)\frac{\partial\mathcal{L}_{\text{g}\leftrightarrow\text{i}}}{\partial\theta_{\text{imu}}}, \tag{2}$$

$$\frac{\partial\mathcal{L}^{\text{joint}}_{\text{g}\leftrightarrow 3}}{\partial\theta_{\text{gps}}} = (1-\alpha)\frac{\partial\mathcal{L}_{\text{g}\leftrightarrow\text{i}}}{\partial\theta_{\text{gps}}}. \tag{3}$$

The same derivation applies to the Wi-Fi path.

Accordingly, when learning the GPS encoder via chained modal transfer learning, data from modalities other than the adjacent IMU modality are not used directly, whereas the parameters of the IMU encoder are influenced by both the 3D skeleton and GPS modalities. Consequently, the GPS embedding space is shaped by the 3D skeleton data through the IMU encoder, effectively transferring knowledge from the intermediate modality (3D skeleton) to the indirect modality (GPS). In the appendices, using the information bottleneck method (Tishby et al., 1999), we show that chained modal transfer learning can effectively mitigate the accumulation of embedding errors caused by chained modal alignment (*cf.* Appendix B.1.1).

#### 3.3.1 Fine-tuned Chained Modal Alignment

Furthermore, to mitigate the accumulating embedding errors without sacrificing the flexibility of modality-pair selection, we propose a modal alignment method with a few-shot strategy that leverages a small number of annotated data pairs from each target modality sample and its textual description. We modify the contrastive loss to align each target modality b to modality a as follows:

$$\tilde{\mathcal{L}}_{\text{a}\leftrightarrow\text{b}} = \gamma\mathcal{L}_{\text{a}\leftrightarrow\text{b}} + (1-\gamma)\mathcal{L}_{\text{l}\leftrightarrow\text{b}}. \tag{4}$$

Here, $\mathcal{L}_{\text{l}\leftrightarrow\text{b}}$ denotes the contrastive loss between the few-shot samples of modality b and language, and $\gamma$ is the weighting coefficient.

## 4 Experiments

### 4.1 Experimental Settings

This section outlines the experimental procedure in the following order: datasets, evaluation protocols,

and implementation details. For additional experimental details, including dataset splits and hyperparameter settings, see Appendices C.1 and C.2.

#### 4.1.1 Datasets

In the proposed method, five types of modal alignments were performed across six modalities. Each encoder was used in three stages: contrastive learning (`Train`), model performance validation during training (`Val`), and model effectiveness evaluation (`Test`). To prevent data leakage and accurately assess the effectiveness of the proposed method, we divided three datasets—BABEL, Nymeria, and WiPose—covering all six modalities into subsets for these different combinations.

**BABEL.** This dataset (Punnakkal et al., 2021), which is built on the AMASS dataset (Mahmood et al., 2019), includes human motion data from over 300 participants. Human skeleton and shape were represented using the SMPL model (Loper et al., 2015). Each scene was annotated with a sentence describing the action. The 2D skeleton joints were those common to the 3D skeleton of this dataset and WiPose, whereas the 3D skeleton joints were those common to this dataset and Nymeria.

**Nymeria.** This dataset (Ma et al., 2024) captures human motion from 264 participants. The human skeleton and shape were captured using a motion-capture system with 17 IMUs. Although 310,500 sentences with three granularity levels were annotated per scene, action labels were not provided. Therefore, we assigned action labels to each scene by converting the annotated sentences using ChatGPT (OpenAI et al., 2024). The 3D skeleton joints were defined identically to those in BABEL. For IMU data, we combined the data measured by attaching two IMUs to the wrists.

**WiPose.** This dataset (Zhou et al., 2023) was developed to promote research on estimating human 2D skeletons from Wi-Fi CSI data. It includes 166,600 packets collected from 12 participants. Each packet contains 2D joint coordinates estimated from camera images using AlphaPose (Fang et al., 2022) and CSI data captured via a 5 GHz-band wireless router and Wi-Fi receiver. The 2D skeleton joint definitions follow those used in the BABEL dataset.

**Dataset Splitting Strategy.** To accurately assess zero-shot action recognition accuracy for each modality, we randomly split the datasets, ensuring that action labels in the `Train` and `Val` splits are disjoint from those in the `Test` split. For improved interpretability of cross-modality comparisons, we selected action labels redundantly present across the three aforementioned datasets and designated them as the target unseen actions for emergent zero-shot action recognition. Within each dataset's predefined evaluation subsets, scenes with unseen action classes were assigned to the `Test` split. In the proposed method, to ensure that the datasets required for successive modal alignments can be prepared independently, we divided the BABEL and Nymeria datasets into `BABEL-1`, `BABEL-2`, `Nymeria-1`, and `Nymeria-2` during training (*cf.* Tab. 9).

#### 4.1.2 Evaluation Settings

**Baselines.** Since no existing method directly targets our modal-alignment problem setting, we adopt representative modal-alignment methods as baselines. Fig. 1 illustrates the relationship between the proposed and baseline methods.

*Cross-Modal Multihop Method (Kim et al., 2019, 2024; Berman et al., 2025)*: This *pivot* baseline successively translates the source signal into a modality understood by the downstream model. We train IMUPoser (Mollyn et al., 2023) to lift GPS → IMU → 3D skeletons and Wi-Fi → 2D → 3D skeletons and apply MotionCLIP (Tevet et al., 2022) for zero-shot action recognition. All lifting networks use the same data splits as our method to ensure a fair comparison.

*Direct Modal Alignment (Girdhar et al., 2023; Moon et al., 2023; Zhou et al., 2025)*: Each target modality is directly paired with an intermediate modality, avoiding the accumulation of embedding errors. However, this approach reduces flexibility as paired data must be collected for every target–intermediate combination.

*CLIP-Based Method (Radford et al., 2021; Zhang et al., 2025)*: All target modalities are aligned to language embeddings through CLIP. While this method achieves strong performance, it requires exhaustive text annotations for every training sample, which is labor-intensive.

**Evaluation Metrics.** For emergent zero-shot action recognition, we used macro accuracy and macro F1 (Opitz and Burst, 2021) as evaluation metrics. By treating each action class as a binary classification problem, the accuracy and F1 score for each class were computed and averaged across all classes. This procedure mitigates class

imbalance effects on the evaluation and fixes the chance rate for action classification at 50% accuracy, thereby enhancing interpretability. For the accuracy of known actions in the `Val` split (Tab. 2), we adopted the multiclass action classification metrics from a prior study (Tevet et al., 2022).

### 4.2 Implementation Details

To ensure fair evaluation, all methods employed the same encoder architecture as the 3D skeleton encoder of MotionCLIP (Tevet et al., 2022) for modal alignment. For the language encoder $f_{\text{lang}}$, we utilized the publicly available MotionCLIP model. The modality-agnostic embedding $\mathbf{z}$ (Sec. 3.2.1) was set to a dimensionality of 512. As the data pairs used for each modal alignment are employed to compute embedding pairs, strict temporal synchronization between inputs is not required. For the cross-modal multihop baseline, we adopted the IMUPoser (Mollyn et al., 2023) architecture—optimized for multidimensional regression of temporal data—for the IMU, GPS, 2D skeleton, and Wi-Fi modalities. A coarse-to-fine hyperparameter search was conducted on the `Val` split to optimize all methods (*cf.* Tab. 10). At inference, following prior work (Girdhar et al., 2023), we encode the text query and each modality's input into embeddings and compute similarity scores using cosine similarity.

### 4.3 Chained Modal Alignment

#### 4.3.1 Comparison with Baselines

Tab. 1 summarizes the emergent zero-shot action recognition accuracy of the proposed chained modal alignment and baselines.[4] Tab. 2 compares accuracies for known classes on the `Val` split, where training and validation use the same dataset, for the proposed method and the cross-modal multihop baseline. Across Tabs. 1 and 2, the proposed method outperforms the cross-modal multihop baseline by several to over ten percentage points in macro accuracy, macro F1, and overall multiclass accuracy. Furthermore, the proposed method achieves accuracy comparable to the direct modal alignment baseline—trained under the favorable condition of error-free alignment—and is within a few percentage points of the CLIP-based method, which relies on textual annotations for every alignment. These results demonstrate that the proposed method can generate modality-agnostic embeddings for indirect modalities without requiring text annotations or direct pairings with the intermediate modality.

[4]As different modalities may be evaluated on different datasets, direct accuracy comparisons across modalities are not meaningful.

Table 1: Comparison of emergent zero-shot action-recognition performance[4] (%) for chained modal alignment versus baselines. Gray numbers indicate settings inherently more favorable than ours—direct intermediate-modality or text-annotated alignment (*cf.* Fig. 1).

| Eval. | Modality / Test set | 3D Skel. (BABEL) | 2D Skel. (BABEL) | Wi-Fi (WiPose) | IMU (Nymeria) | GPS (Nymeria) | Avg. |
|---|---|---|---|---|---|---|---|
| Acc. (%) | Multihop | 76.9 | 75.9 | 64.5 | 66.2 | 62.3 | 69.1 |
| | Ours | 76.9 | 75.4 | 67.7 | 71.0 | 70.0 | **72.2** |
| | Direct | 76.9 | 75.4 | 68.4 | 71.0 | 69.3 | 72.2 |
| | CLIP | 76.9 | 75.7 | 72.3 | 72.1 | 70.6 | 73.5 |
| F1. (%) | Multihop | 77.0 | 75.9 | 63.6 | 66.0 | 60.2 | 68.5 |
| | Ours | 77.0 | 73.0 | 67.1 | 70.5 | 68.0 | **71.1** |
| | Direct | 77.0 | 73.0 | 67.8 | 70.5 | 67.4 | 71.1 |
| | CLIP | 77.0 | 72.6 | 72.1 | 71.1 | 70.4 | 72.6 |

Table 2: Comparison of recognition accuracy[4] (%) for known action classes between chained modal alignment and baselines.

| Modality / Test set | 3D Skel. (BABEL) | 2D Skel. (BABEL) | Wi-Fi (WiPose) | IMU (Nymeria) | GPS (Nymeria) | Avg. |
|---|---|---|---|---|---|---|
| Multihop | 33.8 | 33.3 | 14.9 | 21.7 | 15.7 | 23.9 |
| Ours | 33.8 | 30.6 | 15.4 | 33.7 | 28.6 | **28.4** |

#### 4.3.2 Robustness Evaluation

In the emergent zero-shot action-recognition experiment, we injected Gaussian noise into GPS and Wi-Fi inputs and randomly masked frames to simulate missing data. We then evaluated the accuracies of the chained modal alignment and the cross-modal multihop baseline. Fig. 4 presents the comparison. The accuracy gap between our method and the baseline widens as noise levels and missing frames increase. This finding shows that our method avoids the error accumulation observed in the cross-modal multihop baseline, caused by repeated modality-to-modality conversions at inference time, thereby yielding a more robust representation.

### 4.4 Chained Modal Transfer Learning

Building on the chained modal alignment experiments, we next investigated whether knowledge on unseen classes, added only to 3D skeleton–language alignment data, can be transferred to indirect-modality encoders through chained modal transfer learning. Specifically, we added scenes with unseen classes solely to the training data for 3D skeleton–language alignment (`BABEL-1-All_Actions` in Tab. 9), maintaining the same set of action classes for all other modal

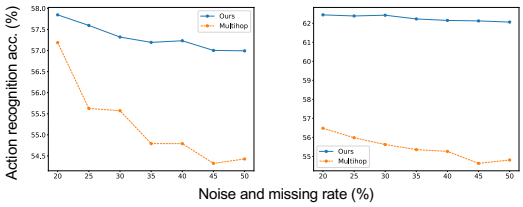



Figure 4: Comparison of the robustness of chained modal alignment and the cross-modal multihop baseline for the Wi-Fi (left) and GPS (right) modalities.

Table 3: Comparison of emergent zero-shot action recognition accuracy (%) between chained modal alignment and chained modal transfer learning.

| Samples (%) | w/ transfer | Wi-Fi | GPS | Avg. |
|---|---|---|---|---|
| 50 | | 62.8 | 67.6 | 65.2 |
| | ✓ | 66.3 | 67.7 | **67.0** |
| 10 | | 57.2 | 66.4 | 61.8 |
| | ✓ | 60.0 | 67.5 | **63.7** |

Table 4: Fine-tuning results (% accuracy) using annotated data for chained modal alignment. Gray numbers indicate settings inherently more favorable than ours—text-annotated alignment.

| Method | # Text annotations | IMU | GPS | Avg. | Progress (%) |
|---|---|---|---|---|---|
| w/o fine-tuning | 0 | 71.0 | 70.0 | 70.5 | 0 |
| w/ fine-tuning | 100 (0.07%) | 71.0 | 70.1 | 70.6 | 6 |
| | 1k (0.7%) | 71.7 | 70.1 | 70.9 | 47 |
| CLIP-based baseline | 138k (100%) | 72.1 | 70.6 | 71.4 | 100 |

alignments and gradually reducing the number of training samples in the indirect modalities to assess the impact of transfer learning. Notably, the training data for chained modal alignment and chained modal transfer learning are identical for the target modalities.

Tab. 3 compares the emergent zero-shot action recognition accuracy of chained modal alignment with that of chained modal transfer learning under these conditions. Chained modal transfer learning outperformed chained modal alignment (where each modal alignment is learned independently) by several percentage points on average. Therefore, chained modal transfer learning can indirectly transfer knowledge learned by an intermediate modality to the encoder of an indirect modality via another modality, thereby improving generalization.

### 4.5 Fine-Tuned Chained Modal Alignment

Tab. 4 presents the emergent zero-shot recognition accuracy obtained by applying the fine-tuned chained modal alignment, which incorporates annotated sentences during training. The results demonstrate that adding few-shot samples brings the accuracy of the baseline chained modal alignment close to that of the CLIP-based baseline, which directly aligns all modalities with language using annotated data. Therefore, in applications where limited manual annotation is feasible, the proposed method mitigates the alignment error to language for indirect modalities while preserving flexibility in modality-pair selection.

Table 5: Ablation study on direct vs. chained alignment for GPS and Wi-Fi (% accuracy).

| Method | Wi-Fi | GPS | Avg. |
|---|---|---|---|
| Direct | 68.4 | 69.3 | **68.9** |
| Chain | 67.7 | 70.0 | **68.9** |

Table 6: Ablation study on chained modal alignment between two action groups with varying planar displacements (% accuracy).

| Action group | IMU | GPS | GPS-IMU |
|---|---|---|---|
| Planar | 74.7 | 77.4 | **+2.7** pp |
| Non-planar | 61.1 | 58.3 | -2.8 pp |

### 4.6 In-depth Analysis

#### 4.6.1 Accumulated Embedding Errors

This section evaluates the issue of accumulated embedding errors in chained modal alignment discussed in Appendix B.1.1. First, Tab. 5 presents the emergent zero-shot action recognition accuracy when the indirect modalities GPS and Wi-Fi are directly aligned with the intermediate 3D skeleton, *i.e.*, direct modal alignment. No performance drop in average accuracy is observed under chained modal alignment compared to direct alignment. Furthermore, we divided the target actions into two groups with different displacement magnitudes on the plane, and Tab. 6 reports the accuracy using chained modal alignment for the IMU and GPS modalities. For the indirect modality, GPS, the proposed method achieves higher accuracy than for the directly aligned modality, IMU, on the action group involving planar movement. Overall, these findings suggest that while the learned embedding space is constrained by the aligned sensors' characteristics (see Appendix C.3.1), surprisingly, error accumulation appears to be negligible in practice under our proposed setting, and the accuracy does not deteriorate.

#### 4.6.2 Evaluation of Language Embeddings

Tab. 7 reports the impact of language embeddings on our method by (i) varying the expressive power of the language encoder used in chained modal alignment and (ii) rephrasing query sentences supplied for emergent zero-shot action recognition. Specifically, we replaced the MotionCLIP-based language encoder $f_{\text{lang}}$ described in Sec. 3.2.1 with

Table 7: Ablation study of the language encoder and query formulations on emergent zero-shot action recognition accuracy (%).

| Query | Text Enc. | IMU | GPS | Avg. |
|---|---|---|---|---|
| word | Mo.CLIP | 71.0 | 70.0 | **70.5** |
| | ChatGPT | 68.5 | 69.0 | 68.8 |
| short | Mo.CLIP | 70.5 | 69.7 | 70.1 |
| | ChatGPT | 70.8 | 70.7 | **70.8** |
| long | Mo.CLIP | 68.9 | 67.4 | 68.2 |
| | ChatGPT | 72.0 | 68.6 | **70.3** |

Table 8: Examples of paraphrased query sentences describing target actions.

| Query | Example |
|---|---|
| word | walk |
| short | Human action of walk |
| long | To walk, a person stands upright and shifts weight from one foot to the other in a rhythmic motion, .... |

ChatGPT (OpenAI et al., 2024) without fine-tuning. In addition, as shown in Tab. 8, each query sentence was paraphrased into three variants using ChatGPT. Based on Tab. 7, enhancing the expressive power of the language encoder with ChatGPT improves accuracy when target actions are described in `short` or `long` sentences, compared to using only single-`word` queries.

### 4.6.3 Choice of Intermediate Modality

We further examined whether the proposed method critically depends on using 3D skeletons as the intermediate modality. We instead treated GPS as the intermediate modality and directly aligned IMU with GPS, while keeping the other protocols unchanged. The IMU accuracy remained comparable and slightly improved (71.0% vs. 73.8%), suggesting that the intermediate modality can be flexibly chosen when it is tightly coupled to the target modality.

## 5 Conclusion

This study proposed a cross-modal learning framework that embeds data from six modalities—language, temporal 2D and 3D skeletons, IMU, GPS, and Wi-Fi—into a shared, modality-agnostic embedding space. Our primary contribution, chained modal alignment, sequentially aligns these modalities via temporal 3D skeletons, enabling emergent zero-shot action recognition within the shared space. Experiments on three public datasets spanning all six modalities quantitatively validated the effectiveness of our method.

## 6 Limitations

This paper aims to advance machine learning by enabling scalable zero-shot action recognition across heterogeneous sensing modalities through chained modal alignment. By reducing the need for large-scale text-paired datasets for every modality, the proposed framework may benefit applications such as activity understanding, rehabilitation, and assistive technologies, particularly in settings where cameras are undesirable or impractical.

At the same time, the modalities considered in this work (*e.g.*, Wi-Fi CSI, IMU, and GPS-derived signals) can raise privacy and fairness concerns. Even without identifiable imagery, motion-related signals may allow re-identification from gait or movement patterns, facilitate inference of sensitive attributes, and expose sensitive location traces. Moreover, the same techniques could be misused for covert surveillance, occupancy tracking, or behavioral inference without informed consent. We oppose and do not condone such uses.

Our study relies exclusively on established, publicly available datasets (BABEL, Nymeria, WiPose); we collected no new human-subject data and comply with each dataset's license and terms of use. We do not attempt to re-identify individuals or infer protected attributes. We also acknowledge that we did not conduct a dedicated analysis of demographic biases or representativeness across age groups, body types, or mobility impairments. A comprehensive assessment and mitigation of these risks—including dataset curation, auditing, and policy-compliant data handling—are important directions for future work.

## References


Sungtae An, Alessio Medda, Michael N. Sawka, Clayton J. Hutto, Mindy L. Millard-Stafford, Scott Appling, Kristine L. S. Richardson, and Omer T. Inan. 2021. Adaptnet: Human activity recognition via bilateral domain adaptation using semi-supervised deep translation networks. *IEEE Sensors Journal*, 21(18):20398–20411.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. 2025. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *PAMI*, 41(2):423–443.

Nimrod Berman, Omkar Joglekar, Eitan Kosman, Dotan Di Castro, and Omri Azencot. 2025. Towards general modality translation with contrastive and predictive latent diffusion bridge. In *NeurIPS*.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang,

Xizhou Zhu, Lewei Lu, et al. 2024. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. 2022. AlphaPose: Whole-body regional multi-person pose estimation and tracking in real-time. *PAMI*.

Gemini Team. 2025. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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. In *CVPR*.

GitHub, Inc. GitHub Copilot.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. 2024. OneLLM: One framework to align all modalities with language. In *CVPR*.

Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Curtis Hawthorne, Andrew M Dai, Matthew D Hoffman, and Douglas Eck. 2018. Music transformer: Generating music with long-term structure. *arXiv preprint arXiv:1809.04281*.

Seunghee Kim, Changhyeon Kim, and Taeuk Kim. 2024. Fcmr: Robust evaluation of financial cross-modal multi-hop reasoning. *arXiv preprint 2412.12567*.

Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based transfer learning for neural machine translation between non-English languages. In *EMNLP-IJCNLP*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NeurIPS*.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16.

Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, Kevin Bailey, David Soriano Fosas, C. Karen Liu, Ziwei Liu, Jakob Engel, Renzo De Nardi, and Richard Newcombe. 2024. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *ECCV*.

Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. 2022. X-CLIP: End-to-end multi-grained contrastive learning for video-text retrieval. In *ACMMM*.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. 2019. Amass: Archive of motion capture as surface shapes. In *ICCV*.

Vimal Mollyn, Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja. 2023. IMUPoser: Full-body pose estimation using imus in phones, watches, and earbuds. In *CHI*.

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Aparajita Saraf, Amy Bearman, and Babak Damavandi. 2023. IMU2CLIP: Language-grounded motion sensor translation with multimodal contrastive learning. In *EMNLP*.

Shuteng Niu, Yongxin Liu, Jian Wang, and Houbing Song. 2020. A decade survey of transfer learning (2010–2020). *IEEE Transactions on Artificial Intelligence*, 1(2):151–166.

OpenAI. ChatGPT.

OpenAI, Josh Achiam, Steven Adler, et al. 2024. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Juri Opitz and Sebastian Burst. 2021. Macro f1 and macro f1. *arXiv preprint arXiv:1911.03347*.

Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. 2021. BABEL: Bodies, action and behavior with english labels. In *CVPR*.

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. In *ICML*.

Runway AI, Inc. Gen-4.

Fumiaki Sato, Ryo Hachiuma, and Taiki Sekii. 2023. Prompt-guided zero-shot anomaly action recognition using pretrained deep skeleton features. In *CVPR*.

Sofia Scataglini, Eveline Abts, Cas Van Bocxlaer, Maxime Van den Bussche, Sara Meletani, and Steven Truijen. 2024. Accuracy, validity, and reliability of markerless camera-based 3d motion capture systems versus marker-based 3d motion capture systems in gait analysis: A systematic review and meta-analysis. *Sensors*, 24(11).

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *arXiv preprint arXiv:1212.0402*.

Kun Su, Xiulong Liu, and Eli Shlizerman. 2020. PREDICT & CLUSTER: Unsupervised skeleton based action recognition. In *CVPR*.

Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. 2022. MotionCLIP: Exposing Human Motion Generation to CLIP Space. In *ECCV*.

Naftali Tishby, Fernando C. Pereira, and William Bialek. 1999. The information bottleneck method. In *The 37th annual Allerton Conference on Communication, Control, and Computing*.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pretraining. In *NeurIPS*.

Hugo Touvron, Louis Martin, Kevin Stone, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. 2023. GeoCLIP: Clip-inspired alignment between locations and images for effective worldwide geo-localization. In *NeurIPS*.

Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. 2017. Untrimmednets for weakly supervised action recognition and detection. In *CVPR*.

Mengmeng Wang, Jiazheng Xing, and Yong Liu. 2022. ActionCLIP: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*.

Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, and Heng Tao Shen. 2024. Cross-modal retrieval: A systematic review of methods and future directions. *Proceedings of the IEEE*, 112(11):1716–1754.

Fei Wu, Qingzhong Wang, Jiang Bian, Ning Ding, Feixiang Lu, Jun Cheng, Dejing Dou, and Haoyi Xiong. 2023. A survey on video action recognition in sports: Datasets, methods and applications. *IEEE Transactions on Multimedia*, 25:7943–7966.

Shaoxuan Xu, Menglu Cui, Chengxiang Huang, Hongfa Wang, and Di Hu. 2025. Balancebenchmark: A survey for multimodal imbalance learning. *arXiv preprint arXiv:2502.10816*.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yugang Jiang, and Xipeng Qiu. 2024. AnyGPT: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*.

Yuwei Zhang, Kumar Ayush, Siyuan Qiao, A. Ali Heydari, Girish Narayanswamy, Maxwell A. Xu, Ahmed A. Metwally, Shawn Xu, Jake Garrison, Xuhai Xu, Tim Althoff, Yun Liu, Pushmeet Kohli, Jiening Zhan, Mark Malhotra, Shwetak Patel, Cecilia Mascolo, Xin Liu, Daniel McDuff, and Yuzhe Yang. 2025. SensorLM: Learning the language of wearable sensors. In *NeurIPS*.

Bo Zhou, Liulei Li, Yujia Wang, Huafeng Liu, Yazhou Yao, and Wenguan Wang. 2025. UNIALIGN: Scaling multimodal alignment within one unified model. In *CVPR*.

Yue Zhou, Caojie Xu, Lu Zhao, Aichun Zhu, Fangqiang Hu, and Yifeng Li. 2023. Csi-former: Pay more attention to pose estimation with wifi. *Entropy*, 25(1).

# A Related Work

This section reviews contrastive learning-based zero-shot action recognition, the task addressed in this study. We also discuss unsupervised cross-modal transfer learning. For a comprehensive survey on cross-modal learning, see (Baltrušaitis et al., 2019; Wang et al., 2024).

## A.1 Zero-Shot Action Recognition via Modal Alignment

Since AlexNet (Krizhevsky et al., 2012) emerged in 2012, the field has witnessed a leap in action-recognition performance for video modalities through the advent of deep learning. Many studies have explored video action recognition in various settings, including supervised learning (Soomro et al., 2012; Carreira and Zisserman, 2017), weakly supervised learning (Wang et al., 2017), and unsupervised learning (Su et al., 2020). In parallel with advancements in natural language processing, CLIP (Radford et al., 2021) was introduced to embed the image modality into a shared feature space via contrastive learning and has since been applied to video-based action recognition (Wang et al., 2022; Ma et al., 2022).

The concept of contrastive learning in CLIP has been extended to modalities beyond images and videos. For instance, MotionCLIP (Tevet et al., 2022) utilized the CLIP framework for the 3D skeleton modality, enabling zero-shot action recognition and the generation of 3D skeleton sequences from text describing human actions. Sato et al. (2023) introduced CLIP into the 2D skeleton modality for zero-shot and anomaly-action recognition. IMU2CLIP (Moon et al., 2023) applied a similar approach for retrieval and action-recognition tasks with IMU data. GeoCLIP (Vivanco Cepeda et al., 2023) utilized paired GPS data and images for training, leveraging a prealigned image encoder with language to enable text-query-based geo-localization. For the Wi-Fi modality, methods have been proposed to reconstruct 2D skeletal poses from CSI obtained via Wi-Fi transmitters and receivers (Zhou et al., 2023). Combining these with MotionCLIP could enable zero-shot action recognition.

## A.2 Unsupervised Cross-Modal Transfer Learning

Cross-modal transfer learning (Niu et al., 2020; An et al., 2021; Tong et al., 2022) aims to transfer knowledge from a well-trained model of one modality to a target modality with limited data or annotations, thereby improving the representation and generalization of the target modality. Recently, unsupervised approaches have enabled cross-modal transfer learning without requiring annotated data or textual labels for the target modalities. Among the studies discussed in Sec. 2.1, those that assume automatic pairing between the target modality and another modality with ample training resources (Girdhar et al., 2023; Moon et al., 2023) are categorized as unsupervised cross-modal transfer learning methods.

Compared with these approaches, our chained modal transfer learning method simultaneously achieves two goals: (1) it indirectly aligns the intermediate and indirect modalities, and (2) it enhances the generalization ability of the DNN for the indirect modality via knowledge transferred from the intermediate modality.

# B Method

## B.1 Training Framework

Fig. 2 illustrates the chaining of modalities, which introduces *indirect modalities* (*e.g.*, GPS, Wi-Fi) that are aligned with the intermediate modality only indirectly via another modality.

Algorithms 1 and 2 present the pseudocode for the chained modal alignment and transfer learning procedures.

### B.1.1 Analysis Based on the Information Bottleneck Method

In this section, we analyze the advantages of chained modal transfer learning over chained modal alignment using the information bottleneck (IB) method. Further details on the IB method can be found in (Tishby et al., 1999). Hereafter, we denote the four modalities from language to the indirect modality (GPS or Wi-Fi) as a, b, c, and d. We assume that $f_{\mathrm{a}}$ and $f_{\mathrm{b}}$ are pretrained via chained modal alignment and focus on the behaviors of $f_{\mathrm{c}}$ and $f_{\mathrm{d}}$, which are optimized during chained modal transfer learning.

**Challenges in Chained Modal Alignment.** Let $I(\cdot;\cdot)$ denote mutual information. Chained modal alignment is equivalent to training $f_{\mathrm{c}}$ by maximizing $I(\mathbf{z}_{\mathrm{c}};\mathbf{z}_{\mathrm{b}})$ and fixing it, followed by training $f_{\mathrm{d}}$ by maximizing $I(\mathbf{z}_{\mathrm{d}};\mathbf{z}_{\mathrm{c}})$. Therefore, for the solution $\left(\bar{f}_{\mathrm{c}}, \bar{f}_{\mathrm{d}}\right)$ obtained by chained modal alignment,

**Algorithm 1** Chained Modal Alignment

**Require:** Encoders $f_{\text{lang}}$, $f_{\text{3d}}$, $f_{\text{imu}}$, $f_{\text{gps}}$, $f_{\text{2d}}$, $f_{\text{wifi}}$
**Require:** Training datasets:

- $\left\{\left(\mathcal{X}_{\text{3d}}, \mathcal{X}_{\text{lang}}\right)\right\}$ for 3D skeleton–language
- $\{(\mathcal{X}_{\text{imu}}, \mathcal{X}_{\text{3d}})\}$ for IMU–3D skeleton
- $\left\{\left(\mathcal{X}_{\text{gps}}, \mathcal{X}_{\text{imu}}\right)\right\}$ for GPS–IMU
- $\{(\mathcal{X}_{\text{2d}}, \mathcal{X}_{\text{3d}})\}$ for 2D–3D skeleton
- $\{(\mathcal{X}_{\text{wifi}}, \mathcal{X}_{\text{2d}})\}$ for Wi-Fi–2D skeleton

**Step 1:** Train $f_{\text{3d}}$ and $f_{\text{lang}}$ by minimizing $\mathcal{L}_{3\leftrightarrow\text{l}}$
**Freeze** $f_{\text{3d}}$
**Step 2:** Train $f_{\text{imu}}$ by minimizing $\mathcal{L}_{\text{i}\leftrightarrow 3}$
**Step 3:** Train $f_{\text{2d}}$ by minimizing $\mathcal{L}_{2\leftrightarrow 3}$
**Freeze** $f_{\text{imu}}$, $f_{\text{2d}}$
**Step 4:** Train $f_{\text{gps}}$ by minimizing $\mathcal{L}_{\text{g}\leftrightarrow\text{i}}$
**Step 5:** Train $f_{\text{wifi}}$ by minimizing $\mathcal{L}_{\text{w}\leftrightarrow 2}$
**return** Encoders $f_{\text{lang}}, f_{\text{3d}}, f_{\text{imu}}, f_{\text{2d}}, f_{\text{gps}}, f_{\text{wifi}}$

the data-processing inequality yields

$$I(\bar{\mathbf{z}}_{\text{d}}; \mathbf{z}_{\text{b}}) \leq I(\bar{\mathbf{z}}_{\text{c}}; \mathbf{z}_{\text{b}}), \tag{5}$$

However, because $f$ is a non-invertible transformation that inevitably discards input information and the modalities differ, equality does not hold in practice. Hence, performing a stepwise modal alignment from the intermediate modality b to the indirect modality d inevitably incurs a loss in $I(\mathbf{z}_{\text{d}}; \mathbf{z}_{\text{b}})$.

**Advantages of Chained Modal Transfer Learning.** Since the contrastive loss provides the negative lower bound on mutual information, we approximate the minimization of Eq. (1) by minimizing the IB objective

$$\begin{aligned}\mathcal{L}^{\text{IB}}_{\text{d}\leftarrow\text{b}} =& I(\mathcal{X}_{\text{c}}; \mathbf{z}_{\text{c}}) + I(\mathcal{X}_{\text{d}}; \mathbf{z}_{\text{d}}) \\ &- \lambda \left\{ I(\mathbf{z}_{\text{c}}; \mathbf{z}_{\text{b}}) + I(\mathbf{z}_{\text{d}}; \mathbf{z}_{\text{c}}) \right\}. \end{aligned} \tag{6}$$

We then assume that the embeddings $\mathbf{z}$ retain only the semantic information of the input data $\mathcal{X}$ and consider the case where the information preservation terms $I(\mathbf{z}_{\text{c}}; \mathbf{z}_{\text{b}})$ and $I(\mathbf{z}_{\text{d}}; \mathbf{z}_{\text{c}})$ dominate, and the compression terms are regarded as relatively small. Under this assumption, $\mathcal{L}^{\text{IB}}_{\text{d}\leftarrow\text{b}}$ can be approximated as

$$\tilde{\mathcal{L}}^{\text{IB}}_{\text{d}\leftarrow\text{b}} \simeq -\lambda \{ I(\mathbf{z}_{\text{c}}; \mathbf{z}_{\text{b}}) + I(\mathbf{z}_{\text{d}}; \mathbf{z}_{\text{c}}) \} + \text{const.} \tag{7}$$

**Algorithm 2** Chained Modal Transfer Learning

**Require:** Encoders $f_{\text{3d}}$, $f_{\text{imu}}$, $f_{\text{2d}}$, $f_{\text{gps}}$, $f_{\text{wifi}}$ initialized via chained modal alignment
**Require:** Training datasets:

- $\{(\mathcal{X}_{\text{imu}}, \mathcal{X}_{\text{3d}})\}$ for IMU–3D skeleton
- $\left\{\left(\mathcal{X}_{\text{gps}}, \mathcal{X}_{\text{imu}}\right)\right\}$ for GPS–IMU
- $\{(\mathcal{X}_{\text{2d}}, \mathcal{X}_{\text{3d}})\}$ for 2D–3D skeleton
- $\{(\mathcal{X}_{\text{wifi}}, \mathcal{X}_{\text{2d}})\}$ for Wi-Fi–2D skeleton

**Freeze** $f_{\text{3d}}$
**Step 1:** Train $f_{\text{imu}}$ and $f_{\text{gps}}$ together by minimizing $\mathcal{L}^{\text{joint}}_{\text{g}\leftrightarrow 3}$
**Step 2:** Train $f_{\text{2d}}$ and $f_{\text{wifi}}$ together by minimizing $\mathcal{L}^{\text{joint}}_{\text{w}\leftrightarrow 3}$
**return** Encoders $f_{\text{imu}}, f_{\text{2d}}, f_{\text{gps}}, f_{\text{wifi}}$

Minimizing $\tilde{\mathcal{L}}^{\text{IB}}_{\text{d}\leftarrow\text{b}}$ is therefore equivalent to solving

$$\max_{f_{\text{c}}, f_{\text{d}}} \left\{ I(\mathbf{z}_{\text{c}}; \mathbf{z}_{\text{b}}) + I(\mathbf{z}_{\text{d}}; \mathbf{z}_{\text{c}}) \right\}. \tag{8}$$

Assuming that the common encoder $f_{\text{c}}$ has sufficient representational capacity, both terms can be increased simultaneously within the range permitted by the data-processing inequality

$$I(\mathbf{z}_{\text{d}}; \mathbf{z}_{\text{b}}) \leq I(\mathbf{z}_{\text{c}}; \mathbf{z}_{\text{b}}). \tag{9}$$

Considering the above, chained modal transfer learning optimizes Eq. (8) within a region that includes the solution $\left(\bar{f}_{\text{c}}, \bar{f}_{\text{d}}\right)$ obtained by chained modal alignment. Therefore, we can obtain the solution $(f^{\star}_{\text{c}}, f^{\star}_{\text{d}})$ that satisfies

$$I(\mathbf{z}^{\star}_{\text{c}}; \mathbf{z}_{\text{b}}) + I(\mathbf{z}^{\star}_{\text{d}}; \mathbf{z}^{\star}_{\text{c}}) \;\geq\; I(\bar{\mathbf{z}}_{\text{c}}; \mathbf{z}_{\text{b}}) + I(\bar{\mathbf{z}}_{\text{d}}; \bar{\mathbf{z}}_{\text{c}}). \tag{10}$$

Consequently, the emergent zero-shot action recognition performance of the indirect modality d is expected to be at least comparable and typically superior when chained modal transfer learning is applied.

# C Experiments

## C.1 Experimental Settings

### C.1.1 Datasets

To prevent data leakage and accurately assess the effectiveness of the proposed method, we divided three datasets—BABEL, Nymeria, and WiPose—covering all six modalities into subsets for these different combinations. Tab. 9 lists the dataset assignments for each modality.

Table 9: Representative dataset assignments in the evaluation experiments.

| Experiment | Wi-Fi–2D Skel. | 2D Skel.–3D Skel. | Lang.–3D Skel. | IMU–3D Skel. | GPS–IMU |
|---|---|---|---|---|---|
| Chained modal alignment<br>Chained modal transfer learning | WiPose | BABEL-2 | BABEL-1<br>BABEL-1-All_Actions | Nymeria-1 | Nymeria-2 |

Table 10: Hyperparameter settings.

| Modality | 3D Skel. | IMU | GPS | 2D Skel. | Wi-Fi |
|---|---|---|---|---|---|
| | | | Baseline | | |
| Learning rate | 1.0e-03 | 1.0e-3 | 1.0e-4 | 1.0e-4 | 1.0e-4 |
| Weight decay | 0.01 | 0.2 | 0.2 | 0.05 | 0.1 |
| Batch size | 20 | 20 | 20 | 20 | 20 |
| Temperature | - | - | - | - | - |
| Total epochs | 120 | 100 | 100 | 100 | 100 |
| | | | Ours | | |
| Learning rate | 1.0e-04 | 1.0e-6 | 2.0e-6 | 1.0e-4 | 1.0e-4 |
| Weight decay | 0.01 | 0.5 | 0.5 | 0.05 | 0.5 |
| Batch size | 20 | 20 | 20 | 20 | 20 |
| Temperature | - | 1.0 | 1.0 | 1.0 | 1.0 |
| Total epochs | 120 | 100 | 100 | 100 | 100 |

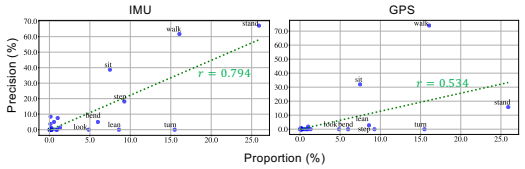



Figure 5: Relationship between the proportion of training samples per action class and per-class precision.

Table 11: Comparison of emergent zero-shot action recognition accuracy for the GPS modality with and without overlapping training samples.

| GPS–IMU Overlap (%) | Dataset in GPS-IMU | GPS Acc. (%) |
|---|---|---|
| 0 | Nymeria-2 | 70.0 |
| 50 | Nymeria-1 + Nymeria-2 | 69.7 |
| 100 | | 69.6 |

## C.2 Implementation Details

Tab. 10 summarizes the hyperparameters employed to train the proposed and baseline methods.

## C.3 In-depth Analysis

### C.3.1 Verification Based on Target Action Characteristics

Fig. 5 illustrates the relationship between the proportion of training samples per known action class and per-class precision for the IMU and GPS modalities using chained modal alignment on the Val split. The correlation between the training-sample proportion and per-class precision decreased from 0.79 for IMU to 0.53 for GPS. As BABEL is primarily designed for 3D skeleton-based action recognition, it is ill-posed for GPS-based action recognition, limiting the set of recognizable actions. This is reflected in the results.

### C.3.2 Training Sample Selection

This section examines whether training samples must overlap between chained modal alignment variants, with a focus on the IMU and GPS modalities. Using the combined dataset, which merges Nymeria-1 and Nymeria-2, we deliberately overlapped the samples learned by the GPS and IMU encoders.

Tab. 11 compares emergent zero-shot action recognition accuracy for chained modal alignment when the proportion of overlapping samples varies at two levels. The results show minimal accuracy differences between settings with and without sample overlap for the IMU–3D skeleton and GPS–IMU alignments. Hence, the proposed method does not require overlapping training samples across adjacent alignments.