

AURA-MFM: 詳細な行動理解のための CLIP を用いた マルチモーダル基盤モデルの構築

松石 康希^{2,a)} 浮田 嵩祐^{2,b)} 大北 剛^{c)}

概要: 近年, ウェアラブルデバイスの普及に伴い, IMU を用いた行動解析が重要性を増している. 医療やロボティクスなど広範な応用が期待される中, 最近の研究ではユニモーダル解析だけでなくマルチモーダル解析への関心が高まっている. いくつかの既存研究では一人称視点動画やテキストなどを導入したマルチモーダル基盤モデルが提案されているが, 人間の全身動作の詳細な解析には至っていない. そこで, 本研究では新たに三人称視点動画とモーションキャプチャを導入し, IMU とテキストを加えた 4 つのモダリティから成る基盤モデルとして, AURA-MFM(**A**ctivity **U**nderstanding and **R**epresentations **A**lignment - **M**ultimodal **F**oundation **M**odel) を提案する. 三人称視点動画とモーションキャプチャデータを用いることで, 一人称視点では捉えきれない人間の全身動作の詳細かつ多角的な解析を行うことを目指す. また, Transformer ベースの IMU エンコーダを採用し, 基盤モデルとしての性能向上を図る. 実験ではリトリバルタスクと行動認識タスクにより評価を行い, 既存手法を上回る性能が得られた. 特に, 行動認識タスクのゼロショット分類では, 既存手法の結果 (F 値: 0.0747, Accuracy: 0.1961) と比較して大幅に上回る分類性能 (F 値: 0.6226, Accuracy: 0.7320) が得られた.

AURA-MFM: Activity Understanding and Representations Alignment - Multimodal Foundation Model

Abstract: In recent years, the widespread adoption of wearable devices has highlighted the growing importance of behavior analysis using IMU. While applications span diverse fields such as healthcare and robotics, recent studies have increasingly focused on multimodal analysis, in addition to unimodal analysis. Some existing research has proposed multimodal foundational models incorporating first-person video and text data; however, these models have yet to achieve detailed analysis of full-body human activity. To address this limitation, we propose AURA-MFM: Activity Understanding and Representations Alignment - Multimodal Foundation Model, a foundational model integrating four modalities: third-person video, motion capture, IMU, and text. By utilizing third-person video and motion capture data, the model enables a detailed and multidimensional understanding of human activity, which is difficult to capture with first-person perspectives alone. Additionally, a Transformer-based IMU encoder is employed to enhance the model's overall performance. Experimental evaluations conducted on retrieval tasks and activity recognition tasks demonstrate that our proposed model outperforms the existing method. Notably, in the zero-shot classification for action recognition, our method achieved significantly higher performance (F1-score: 0.6226, Accuracy: 0.7320) compared to the existing method (F1-score: 0.0747, Accuracy: 0.1961).

1. はじめに

近年, スマートグラスやウェアラブルデバイスの普及に

伴い, 加速度センサやジャイロセンサ (IMU) などのデバイス内蔵センサーを活用した人間の行動解析が注目を集めている. この技術は医療, スポーツ, リハビリテーション, ロボティクス等, 広範な応用が期待されている. 現状では IMU データやビデオなど単一のモダリティに基づく解析が主要な研究テーマとして進展している一方, これらのデータソースを統合してより性能や汎用性を高めたマルチモーダル基盤モデルの構築に向けた試みが増えつつある. この

¹ 九州工業大学大学院情報工学府情報創成工学専攻
680-4, Kawazu, Iizuka, Fukuoka 820-8502, Japan
² 九州工業大学大学院情報工学府情報創成工学専攻
680-4, Kawazu, Iizuka, Fukuoka 820-8502, Japan
^{a)} matsuishiki.kouki587@mail.kyutech.jp
^{b)} ukita.kosuke299@mail.kyutech.jp
^{c)} tsuyoshi@ai.kyutech.ac.jp

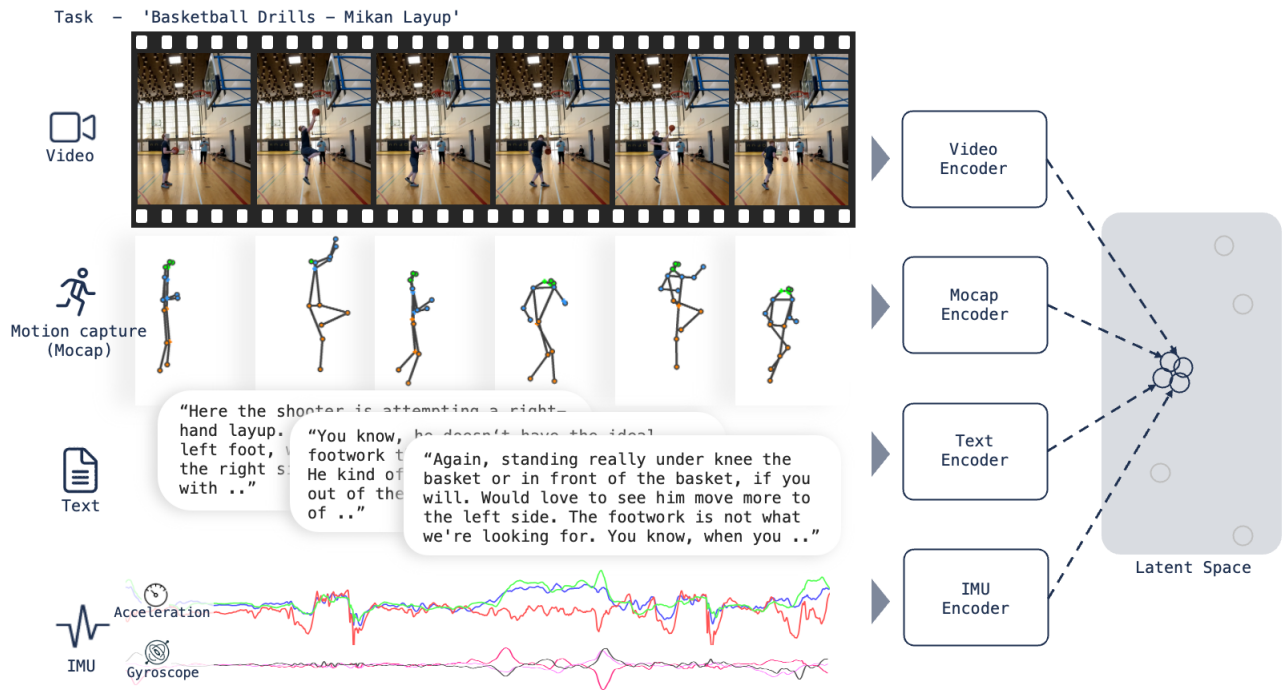


図 1: 4 種類のモダリティによるマルチモーダル基盤モデル (AURA-MFM): 三人称視点動画, モーションキャプチャ, テキスト, IMU (加速度, ジャイロ) の 4 つのモダリティを統合したマルチモーダル基盤モデルの構築を目的とする。

分野では, IMU データを活用した新しいアプローチが次第に注目を集めており, いくつかの有意義な研究成果が報告されている. ([1], [2], [3])

IMU2CLIP[4] は, IMU データをビデオやテキストの表現空間と結び付けるために, RNN (Recurrent Neural Network) ベースの IMU エンコーダを使用した. この手法により, IMU データのみでは認識が困難なアクティビティの性能向上が実現された. また, ImageBind[5] は, 画像を中心とした多モダリティ統合モデルを提案し, タスク一般化において優れた性能を示している. しかし, IMU2CLIP のような RNN ベースのエンコーダは, 時系列データの特性を捉える能力において一定の限界があると考えられる. 特に, 長い時系列データや複雑な動作情報を処理する際, RNN は勾配消失問題や長期依存関係の学習効率の低さといった課題を抱えている.

本研究では, これらの課題を克服するために, Transformer ベースの IMU エンコーダを採用したマルチモーダル基盤モデル「AURA-MFM (Activity Understanding and Representations Alignment - Multimodal Foundation Model)」を提案する. Transformer エンコーダは, 自己注意機構 (self-attention) を活用することで, 長期依存関係を効率的に捉えると同時に, 並列処理による学習効率の向上を可能にする. この特性により, IMU データの複雑な動作情報をより精緻に学習できると考えている.

AURA-MFM は, 図 1 に示す通り, IMU と三人称視点映像, テキスト, さらにはモーションキャプチャを組み合

わせたマルチモーダル基盤モデルを構築することで実現する. 特に, IMU をテキストと画像の混合モデル (CLIP モデル) と接続する IMU2CLIP の研究を基礎としつつ, 三人称視点の動画やテキスト, 人間の体をかたどったモーションキャプチャを統合する新しい枠組みを構築する. 本研究の意義は, 複数のデータモダリティを組み合わせることで, 人間の行動や動作をより精密かつ包括的にモデル化する点にある.

本論文の貢献は以下の 2 点である.

- 三人称視点映像, テキスト, モーションキャプチャ, IMU の 4 つのモダリティを統合したマルチモーダル基盤モデル (AURA-MFM) の構築.
- Transformer に基づいた IMU エンコーダにより, リトリバルタスクおよび行動認識タスクにおいて IMU2CLIP よりも性能を改善する事を示す.

2. 関連研究

IMU2CLIP[4] マルチモーダル学習は, 異なるデータ形式を統合して意味的に豊かな表現を学習する技術として近年注目されている. その一例として, OpenAI の CLIP[6] は画像と言語を結びつける基盤モデルとして広く使用されている一方で, IMU データを CLIP に接続する IMU2CLIP の研究は, IMU データと自然言語, 一人称視点動画を結びつける革新的なアプローチを提供している. この研究は, ウェアラブルデバイスのセンサーデータをテキスト表現や動画表現にマッピングするという新たな可能性を示した

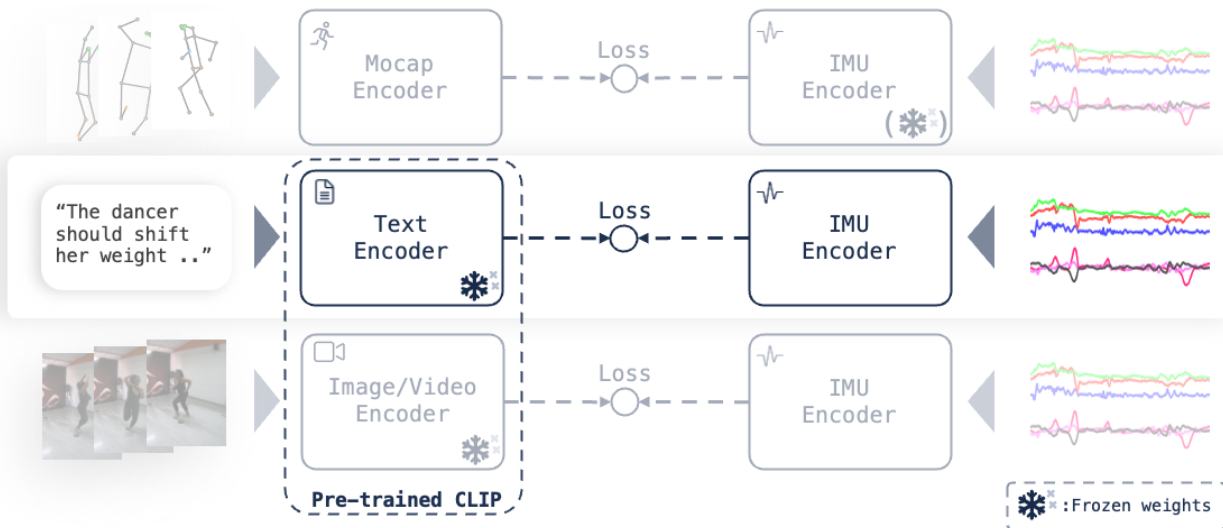


図 2: クロスモーダル対比学習:IMU データと他モダリティによるクロスモーダル対比学習を行う。テキストエンコーダとビデオエンコーダは公開されている訓練済み CLIP のパラメータを使用し、訓練時はパラメータを固定する。図では特にテキストと IMU の訓練を示しているが、ビデオと IMU、モーションキャプチャと IMU、以上 3 種類の訓練により、複数のモダリティを統合した基盤モデルを作成する。

が、三人称視点映像やモーションキャプチャデータの統合には対応していない。

Ego-Exo4D Datasets[7] IMU2CLIP を基盤とした従来のアプローチにはいくつかの課題がある。具体的には、一人称視点動画を用いることにより、カメラ装着者の視点に限られるため、全身の動きや周囲との相互作用を詳細に把握することが難しい点が挙げられる。また、IMU データとの統合も、主に視点が限定されたデータセットを活用していることが原因で、動作解析の範囲が狭まる傾向があると考えられる。これに対し本研究では、三人称視点動画やモーションキャプチャデータを重視し、一人称視点では捉えきれない人間の全身動作を詳細に解析することを目指す。このようなアプローチを可能にするために、我々は Ego-Exo4D データセット [7] を採用する。このデータセットは、一人称視点 (ego) と三人称視点 (exo) の両方を統合的に収集したものであり、人間の行動を多角的に捉える新たな基盤となっている。Ego-Exo4D データセットは、人間の体全体の動きを捉えたモーションキャプチャデータも含まれており、IMU2CLIP や従来の Ego4D データセット [8] では得られなかった視点や特徴を補完する重要な資源であると考えられる。

3. 提案手法

3 節では、我々の提案について述べる。まず、3.1 節では、提案する基盤モデルについて述べる。次に、3.2 節では、我々が新たに構築した IMU エンコーダ、Mocap エンコーダについて述べる。

3.1 提案するマルチモーダル基盤モデル (AURA-MFM)

訓練済み CLIP[4] を使用して IMU エンコーダを対比学習により訓練する手法は IMU2CLIP[4] の画期的な研究成果である。この手法に触発され、我々は IMU に加え、三人称視点動画、モーションキャプチャのための枠組みを新たに導入する事でビデオ、テキスト、モーションキャプチャ、IMU の 4 つのモダリティにより構築される基盤モデル (AURA-MFM) を提案する。図 2 に示す通り、我々が構築するモデルは、IMU と他のモダリティのペアによるクロスモーダル対比学習により実現する。

テキストとビデオのためのエンコーダは訓練済み CLIP のパラメータを固定して使用する。モーションキャプチャと IMU との訓練はテキストまたはビデオで一度訓練された IMU エンコーダを用いてパラメータを固定することで CLIP が表現する潜在空間上にモーションキャプチャと IMU のモダリティを追加することが可能である。それとは別に、CLIP の表現を使用せずにモーションキャプチャと IMU のエンコーダだけで訓練する実験も行う。

クロスモーダル対比学習

使用する損失関数は、IMU2CLIP[4] の調査により、InfoNCE[9]、Triplet[10]、MSE の中で最も性能の高かった InfoNCE を使用する。[4] と同様に、バッチサイズ B の IMU (I) と任意のモダリティ (M) の埋め込みペア: $I(= i_1, i_2, \dots, i_B), M(= m_1, m_2, \dots, m_B)$ を用いて I から M への損失は式 (1)、 M から I への損失は式 (2) で定義される。したがって、最終的な I と M のクロスモーダル損失は、これらの平均で定義される (式 (3))。ここで、 i_i や m_i はエンコードされた埋め込みであり、単位超球面上にある。埋め込みは単位正規化されているため、その類似

性は埋め込み同士の内積により計算できることに注意されたい。

$$L_{I \rightarrow M} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(\langle \mathbf{i}_i, \mathbf{m}_i \rangle)}{\sum_{k=1}^B \exp(\langle \mathbf{i}_i, \mathbf{m}_k \rangle)} \quad (1)$$

$$L_{M \rightarrow I} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(\langle \mathbf{m}_i, \mathbf{i}_i \rangle)}{\sum_{k=1}^B \exp(\langle \mathbf{m}_i, \mathbf{i}_k \rangle)} \quad (2)$$

$$L_{I \leftrightarrow M} = \frac{1}{2} (L_{I \rightarrow M} + L_{M \rightarrow I}) \quad (3)$$

3.2 Transformer に基づいたエンコーダアーキテクチャ

IMU2CLIP[4] では、IMU エンコーダとして、*Stacked 1D-CNN-RNN* を最も性能の高いエンコーダとして提案した。一方で、我々は Transformer に基づいた信号用のアーキテクチャである *SENVt*[11] エンコーダを採用する。*SENVt* からの変更点は2点であり、パッチ埋め込みレイヤのサイズ調整と CLIP と同じ埋め込み次元に調整するためのレイヤを追加することで本実験での IMU エンコーダと Mocap エンコーダとして使用する。図3は、それぞれ上が RNN ベースの *Stacked 1D-CNN-RNN*、下が Transformer ベースの *SENVt* の構造を示している。

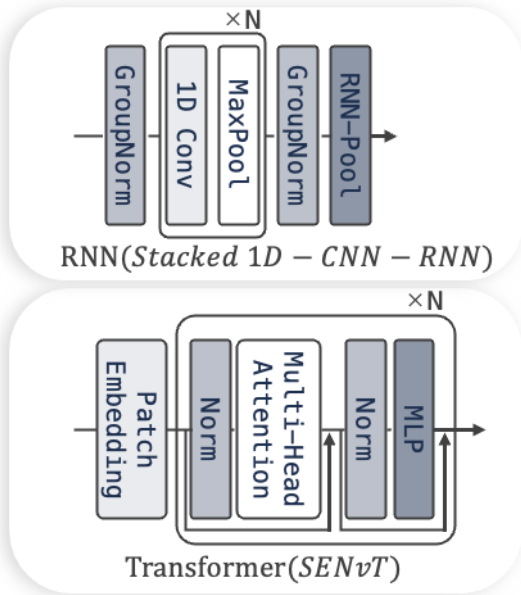


図3: IMU, Mocap のためのエンコーダアーキテクチャ
(上:*Stacked 1D-CNN-RNN*[4], 下:*SENVt*[11])

4. 実験

4 節では、実験に関する前処理と実験結果について述べる。まず、4.1 節では、実験に使用したデータセットと前処理方法について示す。次に、4.2 節では、評価方法であるタスクについて示し、最後に、4.3 節で実験結果を示す。

4.1 データセット

本実験では、Ego-Exo4D データセット [8] を使用する。IMU Aria と呼ばれるスマートグラスの右フレームについている IMU センサにより加速度とジャイロが計測される。それぞれ 1000Hz の解像度で与えられるが、計算コストを考慮して我々は、200Hz にリサンプリングする処理を施す。Text テキストデータは以下に示す EC と AAD に関して実験を行う。

- Expert Commentary(EC): 専門家の解説に加え、アクティビティ名とコメントが付与された秒数をテキストとして追加、さらに使用する CLIP には最大トークン数 77 という制限があるため、77 単語を超える文章は 77 単語までを使用する事で調整する。
- Atomic Action Descriptions(AAD): Ego4d データセットにおけるナレーションと同様の形式であり、非専門家の第三者によって記述された、被験者の行動を説明する短い文章で構成されている。

Video 三人称視点の動画として 4 台の GoPro を使用し同時刻 4 つの視点からの動画が提供されている。4K, 60FPS の動画を我々は、224x224, 10FPS に変換し使用する。スライディングウィンドウは 5 秒に設定する。

Mocap EgoBodyPose と呼称され、右目、左肩、右手首、左足首など計 17 箇所の部位の座標が 10FPS の分解能で提供されている。欠損部位は 0 埋めで補間し、5 秒ごとのスライディングウィンドウを作製する。5 秒以内の欠損区間は線形補間を施す。

表 1: Ego-Exo4D データセット前処理後のデータ数

Data Modality	Train	Validation	Test
IMU $\leftrightarrow$ Mocap (5s)	2045	108	240
IMU $\leftrightarrow$ Video (5s)	121k	16k	1k
IMU $\leftrightarrow$ Text - AAD (5s)	197k	25k	1k
IMU $\leftrightarrow$ Text - EC (5s)	17k	1k	1k
IMU $\rightarrow$ 8 classes (5s)	3385	377	418

4.2 評価方法

図4に示す Task 1 と Task 2 により評価を行う。

Task 1 任意の2つのモダリティに関して、クエリとキーに分け、クエリが与えられた場合にキーのインスタンスを取得するリトリバルタスクを実行する。テストセットで評価され、Recall@k および Mean Reciprocal Rank (MRR) を測定する。

Task 2 IMU ベースの人間行動認識を行う。IMU 信号のスライディングウィンドウが与えられた場合に個別のアクティビティラベルを予測する。使用するアクティビティラベルは、テキストマッチングによるソフト注釈を使用する。テキストによるラベルを使用した分類のため、モーション

表 2: リトリバルタスクの結果

Retrieval tasks	Models	R@1	R@10	R@50	MRR
IMU → Mocap	AURA-MFM RNN	.0250	.1875	.5500	.0849
	Transformer	.2208	.6666	.8541	.3667
Mocap → IMU	AURA-MFM RNN	.0333	.1875	.5291	.0915
	Transformer	.2208	.6541	.8458	.3607
IMU → Text-EC	IMU2CLIP[4]	.0010	.0149	.0719	.0095
	AURA-MFM RNN	.0120	.1030	.4199	.0497
		Transformer	.0199	.1720	.4869
Text-EC → IMU	IMU2CLIP[4]	.0020	.0099	.0780	.0093
	AURA-MFM RNN	.0140	.1180	.4629	.0560
		Transformer	.0309	.1749	.4930
IMU → Text-AAD	IMU2CLIP[4]	.0000	.0041	.0402	.0051
	AURA-MFM RNN	.0140	.1472	.4679	.0620
		Transformer	.0106	.1085	.4185
Text-AAD → IMU	IMU2CLIP[4]	.0008	.0082	.0460	.0063
	AURA-MFM RNN	.0222	.1562	.4745	.0710
		Transformer	.0164	.1233	.4268
IMU → Video	IMU2CLIP[4]	.0000	.0078	.0453	.0057
	AURA-MFM RNN	.0171	.1328	.4367	.0605
		Transformer	.0101	.1109	.4101
Video → IMU	IMU2CLIP[4]	.0000	.0093	.0390	.0053
	AURA-MFM RNN	.0164	.1429	.4734	.0652
		Transformer	.0164	.1140	.4109

キャプチャと IMU 間で作成したエンコーダでの評価は考慮しない。

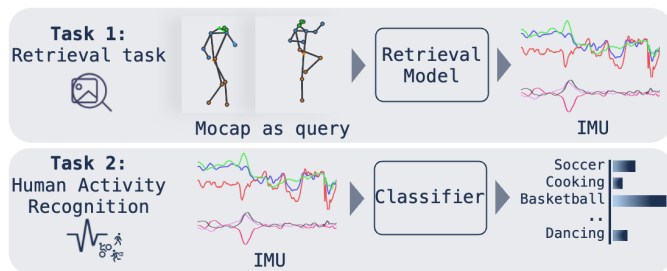


図 4: Task1: 任意のクエリから異なるモダリティへの再現率を測るリトリバルタスク（上）と，Task2:IMU からアクティビティ分類を行う人間行動タスク（下）を示す。

4.3 実験結果

Task1（リトリバルタスク）と，Task2（IMU からの行動認識タスク）について結果と考察を述べる。

リトリバルタスク

表 2 にリトリバルタスクの結果を示す。表中の左に示す IMU→Mocap のなどの記号は，この例では任意の IMU

インスタンスを入力した時の Mocap への Recall 精度を示している。IMU2CLIP[4] が提供する IMU，テキスト，ビデオのモダリティにおいて，我々の提案する AURA-MFM の方が良い性能を示した。最も高いもので，Text - EC → IMU において，MRR は+.0726， $R@\{1, 10, 50\}$ はそれぞれ $\{+.0289, +.1650, +.4150\}$ 向上した事を示す。IMU2CLIP が訓練に使用したデータは Ego4D データセット [8] であり，本実験で我々が使用した Ego-Exo4D データセット [7] とはデータ数やモダリティ，カメラの視点など多少異なり，テストセットとはバイアスがあるため，本来よりも低い結果が出ていることに注意されたい。

次に，我々が新たに導入した Transformer に基づいたエンコーダが Mocap と Text - EC を使用した訓練において性能を改善する事を示す。IMU → Mocap では，RNN(*Stacked 1D-CNN-RNN*) と比較して，MRR は+.2818， $R@\{1, 10, 50\}$ はそれぞれ $\{+.1958, +.4791, +.3041\}$ 向上した。さらに，Text - EC → IMU では，RNN(*Stacked 1D-CNN-RNN*) と比較して，MRR は+.0259， $R@\{1, 10, 50\}$ はそれぞれ $\{+.0169, +.0569, +.0301\}$ 向上した。

図 5 は，IMU→Text - EC と IMU→Mocap のリトリバル結果を示している。図の左の IMU query は，加速度

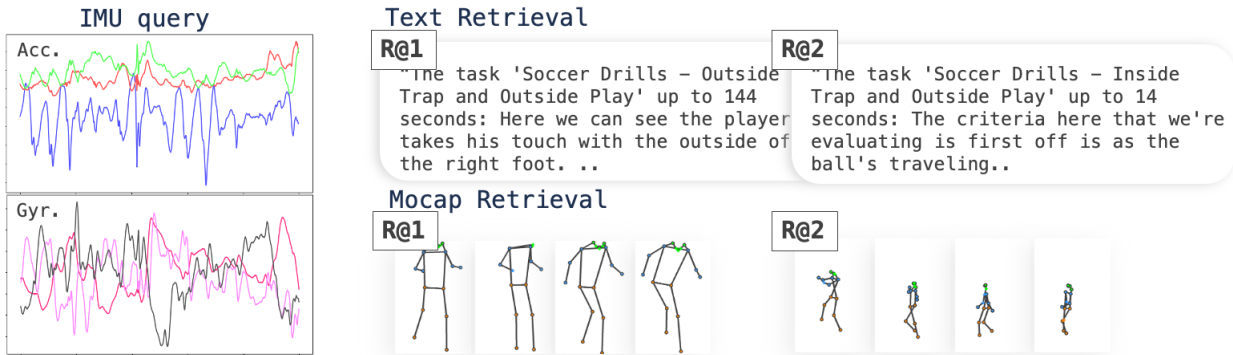


図 5: IMU→ Text-EC と IMU→ Mocap のリトリバル結果

Prompt as text query

"A woman is bouldering indoors,
climbing a colorful artificial rock wall.
She wears a white tank top, striped leggings,
and climbing shoes, using chalk
for grip while reaching for a hold."

Video Retrieval

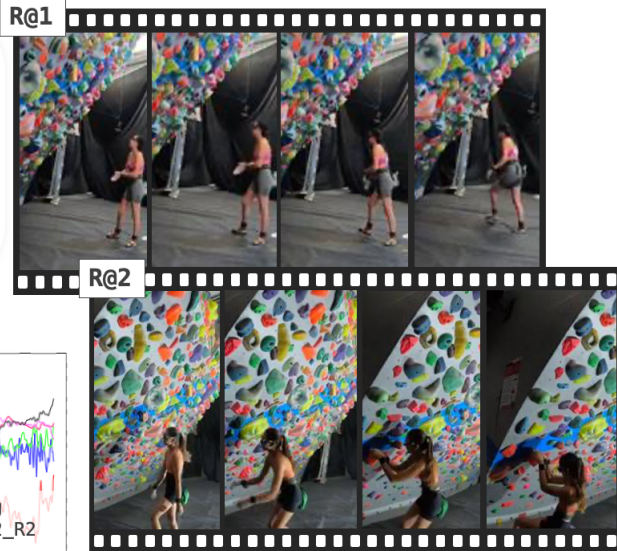


図 6: 自由に作成した Prompt からの IMU と Video のリトリバル結果

とジャイロセンサーの波形であり、この信号のアクティビティ名は、「Soccer Drills - Outside Trap and Outside Play」である。この IMU 信号は、サッカーのパスを受けパスを出す動作を示している。この信号に最もマッチしていると判断されたのが R@1 であり、次の候補が R@2 である。図の右上が Text - EC、右下が Mocap の結果である。まずテキストに関して、R@1、R@2 ともに、サッカーでパスをしているような記述があり、いずれも適していると考え。次に Mocap に関して、R@1 は右足でサッカーボールをトラップしているような動きのように思われ、R@2 は右足でパスを出してサッカーボールを蹴っているような動作に思われる。以上からどちらも適していると考え。図への描画の関係上カメラの視点が変わっているが、使用している Mocap データは 3D データであるため視点は関係ない。

図 6 は、テストセットにないプロンプトテキスト*1を入

力した際の IMU と Video のリトリバル結果を示した図である。“Bouldering”というアクティビティ部分はデータセットに依存するが、自由に作成した文章であり、データセットに含まれないことを確認している。まず IMU に関して、IMU 信号の波形だけでこのアクティビティを理解することは困難であるが、R@1、R@2 ともに付与されているアクティビティ名は「Rock Climbing - LosAndes_V2_R2」であり、“Bouldering”や“Climbing .. rock wall”というテキストから IMU 波形への変換が再現できていると思われる。次に Video に関して、どちらもロッククライミングを表した動画であると言えるが R@1、R@2 ともに開始する直前の動画であり、“is bouldering”という現在進行形の部分や女性の服装などが少し異なると考える。完全な再現はできていないが、このプロンプトを再現できる動画がテストセットにない可能性が高いので、ある程度再現できていると思われる。

IMU からの行動認識タスク

表 3 に行動認識タスクの結果を示す。ゼロショット分類結果は、我々の提案する AURA-MFM の全モデルにおい

*1 訳: “女性がカラフルな人工壁でボルダリングをしている。彼女は白いタンクトップ、ストライプのレギンス、クライミングシューズを身に着け、チョークを使ってグリップを強化しながらホールドに手を伸ばしている。”

表 3: 行動認識タスクの結果

Models		Zeroshot		Transfer learning		Fine-tuning	
		F1	Acc.	F1	Acc.	F1	Acc.
Random Init.	RNN	-	-	-	-	.4078	.6626
	Transformer	-	-	-	-	.6251	.7511
IMU2CLIP[4]		.0747	.1961	.4332	.6315	.5983	.7033
(Text-EC ↔ IMU)	AURA-MFM RNN	.5757	.7057	.5662	.7368	.6228	.7464
	Transformer	.6217	.7081	.6285	.7559	.6491	.7535
(Text-AAD ↔ IMU)	AURA-MFM RNN	.0805	.2009	.3163	.5263	.4801	.6626
	Transformer	.6226	.7320	.7217	.8181	.7506	.8277
(Video↔ IMU)	AURA-MFM RNN	.1111	.2057	.3532	.4808	.5219	.6842
	Transformer	.5901	.7224	.6966	.8062	.7217	.7990

て、IMU2CLIP[4] よりも高い性能を示した。IMU2CLIP[4] の結果 (F1: .0747, Acc: .1961) と比較して、最も高いもので F1: .6226, Acc: .7320 という大幅に向上した結果を得た。

IMU エンコーダアーキテクチャの違いによる結果についても考察する。まず、RNN(*Stacked 1D-CNN-RNN*) で構築したエンコーダは、Text - EC により得られるモデルにおいては高い性能（例: Zeroshot F1:.5757, Acc:.7057）を示したものの、それ以外のものでは大幅な向上は見られなかった。Text - AAD と Video において低くなった要因として、訓練が過学習傾向にあり、評価が不適切である可能性を示唆している。次に、Transformer(*SEnVt*) で構築したエンコーダは、どのチューニングにおいても高い性能を示した。特に、Text - AAD により得られるモデルにおいて、最も高い結果が得られた。こちらは専門家の解説よりも 10 倍ほど多いデータ数で訓練されているため、より頑健な表現獲得につながった可能性があると考ええる。

分類タスクである行動認識において本実験では、ラベルを数値ではなくテキストとして扱い、テキストマッチングによる分類を行っているため、専門家の解説 (Text - EC) や簡単なアクティビティの説明 (Text - AAD) などのテキストによるチューニングを行ったモデルにおいて高い性能を示していると考ええる。

5. 結論

本研究では、IMU2CLIP[4] の研究を基礎としつつ、IMU とテキスト、さらに三人称視点動画や人間の体をかたどったモーションキャプチャを統合する新たなマルチモーダル基盤モデルである AURA-MFM を提案した。一人称視点動画という限られた視点による動画に対し、我々は三人称視点動画やモーションキャプチャデータを重視し一人称視点では捉えきれない人間の全身動作を詳細に解析することを目指した。

実験結果から、行動認識タスクのゼロショット分類結果において我々の提案する AURA-MFM の全モデルにおいて向上することを示し、最も性能良いものでは、IMU2CLIP の結果 (F1: .0747, Acc: .1961) と比較して、F1: .6226, Acc: .7320 という大幅に向上した結果を得た。

本研究では実験結果で示す通り、良好な結果を得ることができたものの、提案手法はまだ研究の初期段階にあり、さらなる性能向上の可能性が残されている。特に、各モダリティ間の統合手法の最適化や複雑な動作の理解に対応するためのアプローチの検討も必要であると考ええる。現状では、1 対 1 のクロスモーダル学習に限定した訓練しか行っていない。今後は、4 つもしくはそれ以上のモダリティを統合し、これらを“同時”に訓練可能なマルチモーダルモデルの構築を目指す。これにより、モダリティ間の相互作用をより効果的に活用し、さらなる性能向上が期待できる。

参考文献

- [1] Sejal Bhalla, Mayank Goel, and Rushil Khurana. Imu2doppler: Cross-modal domain adaptation for doppler-based activity recognition using imu data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 5, No. 4, pp. 1–20, 2021.
- [2] Hyeokhyen Kwon, Catherine Tong, Harish Haresamudram, Yan Gao, Gregory D Abowd, Nicholas D Lane, and Thomas Ploetz. Imutube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 4, No. 3, pp. 1–29, 2020.
- [3] Shohreh Deldari, Hao Xue, Aaqib Saeed, Daniel V Smith, and Flora D Salim. Cocoa: Cross modality contrastive learning for sensor data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 6, No. 3, pp. 1–28, 2022.
- [4] Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Aparajita Saraf, Amy Bearman, and Babak Damavandi. Imu2clip: Language-grounded motion sensor translation with multimodal contrastive learning. In *Findings of the*

Association for Computational Linguistics: EMNLP 2023, pp. 13246–13253, 2023.

- [5] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Man-
nat Singh, Kalyan Vasudev Alwala, Armand Joulin, and
Ishan Misra. Imagebind: One embedding space to bind
them all. In *Proceedings of the IEEE/CVF Confer-
ence on Computer Vision and Pattern Recognition*, pp.
15180–15190, 2023.
- [6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya
Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sas-
try, Amanda Aspell, Pamela Mishkin, Jack Clark, et al.
Learning transferable visual models from natural lan-
guage supervision. In *International conference on ma-
chine learning*, pp. 8748–8763. PMLR, 2021.
- [7] Lorenzo Torresani Kris Kitani Jitendra Malik Triantafyl-
los Afouras Kumar Ashutosh Vijay Baiyya Siddhant
Bansal Bikram Boote et al. Kristen Grauman, An-
drew Westbury. Ego-exo4d: Understanding skilled hu-
man activity from first- and third-person perspectives.
In *Proceedings of the IEEE/CVF Conference on Com-
puter Vision and Pattern Recognition (CVPR)*, pp.
19383–19400, June 2024.
- [8] Eugene Byrne Zachary Chavis Antonino Furnari Rohit
Girdhar Jackson Hamburger Hao Jiang Miao Liu Xingyu
Liu et al. Kristen Grauman, Andrew Westbury. Ego4d:
Around the World in 3,000 Hours of Egocentric Video. In
*IEEE/CVF Computer Vision and Pattern Recognition
(CVPR)*, 2022.
- [9] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Rep-
resentation learning with contrastive predictive coding.
arXiv preprint arXiv:1807.03748, 2018.
- [10] Florian Schroff, Dmitry Kalenichenko, and James
Philbin. Facenet: A unified embedding for face recog-
nition and clustering. In *Proceedings of the IEEE con-
ference on computer vision and pattern recognition*, pp.
815–823, 2015.
- [11] Tsuyoshi Okita, Kosuke Ukita, Koki Matsuishi, Masa-
haru Kagiya, Kodai Hirata, and Asahi Miyazaki. To-
wards llms for sensor data: Multi-task self-supervised
learning. In *Adjunct Proceedings of the 2023 ACM In-
ternational Joint Conference on Pervasive and Ubiqui-
tous Computing & the 2023 ACM International Sym-
posium on Wearable Computing*, pp. 499–504, 2023.