Xingjian's Research Project Summary

1. Research Objective

We aim to develop a comprehensive safety-focused Human Activity Recognition (HAR) system that spans from video-to-IMU data generation to LLM-based activity recognition and reasoning. Our system targets IMU signals from wrist-mounted devices, including smartwatches, wristbands, and phones held in hand. Our key project goals are as follows: ① We aim to introduce a **novel video-to-IMU generation pipeline that addresses the scarcity of large-scale annotated IMU datasets**. While existing SOTA methods primarily target routine daily activities, our approach focuses on synthesizing realistic IMU signals for rare and safety-critical behaviors such as falling, losing balance, fighting, shoving, dragging, jumping from height, and fainting. ② (*Stretch Goal*) We aim to develop a unified, instruction-tuned large language model that takes tokenized IMU sequences and human activity video as input, and performs both **activity classification and fine-grained safety reasoning**.

2. Motivation

① **Limited Coverage of Safety-Critical Behaviors:** Prior works such as IMUGPT 2.0 [5] and Video2IMU [3] have explored IMU signal simulation, but the generated data lacks sufficient coverage of dangerous or abnormal human behaviors. While IMUGPT 2.0 improves over its predecessor via motion filtering and early stopping, it still fails to support a wide range of safety-related scenarios, and the scale and precision of generated sequences are inadequate for fine-grained HAR tasks.

② **Absence of Safety-Level Supervision:** Most existing datasets are designed for general activity classification and lack explicit annotations for safety status (e.g., safe vs. unsafe), making them unsuitable for training models with reasoning capabilities. This hinders the ability of large models to detect subtle differences between semantically similar actions with distinct safety implications.

③ **The Ongoing Need for Synthetic IMU Data Generation:** Synthetic IMU data generation remains essential for HAR, as collecting real-world sensor data is often prohibitively costly to annotate and fundamentally infeasible in scenarios involving rare, dangerous, or privacy-sensitive activities.

- **Data scarcity in IMU-based HAR:** IMU-based Human Activity Recognition (HAR) requires large-scale labeled sensor data, but collecting and labeling IMU signals is costly and labor-intensive, as they are not easily interpretable by human annotators [3, 10].
- **Abundance of easily labeled video data:** Human activity videos are widely available and often easier to annotate via visual inspection or metadata, offering a scalable alternative data source for HAR model training [3, 7].
- **Use of videos to generate synthetic IMU data:** Prior efforts have explored pose-to-sensor mappings using regression or augmentation [6, 9, 11], but Video2IMU advances this by learning to generate IMU-like signals from monocular videos using neural networks [3]. However, its method still presents important limitations in sensor diversity, physical realism, and temporal coherence, especially for complex or safety-critical actions. (Please refer to Table 1)
- **Improved scalability and generalization:** By leveraging synthetic signals from video, the approach reduces annotation cost and supports training models that generalize comparably to those using real IMU data [3].

3. Related Work

Cross-Modality IMU Generation: To overcome limited IMU data, recent works generate virtual sensor signals from text and vision modalities. **Text-based approaches** (e.g., IMUGPT [4] and IMUGPT 2.0 [5]) employ large language models (LLMs) and text-to-motion decoders to synthesize IMU sequences from natural language descriptions. These methods can cover a wide semantic range of activities, *but they lack visual grounding and often produce unrealistic signals for complex physical behaviors.* In contrast, **vision-based approaches** leverage abundant videos to create IMU data. Early pipelines such as IMUTube [2, 7] convert exocentric (third-person) videos into on-body accelerometer streams by extracting 2D/3D poses and applying biomechanical models. Subsequent learning-based methods like Video2IMU [3] and Vi2IMU [8] improve this process using

Table 1 Comparison of virtual IMU generation methods and their support for safety-critical modeling.

Method	Source Modality	Sensor Position	Dangerous Action	Safety Label
IMUTube [2]	Monocular video (2D pose)	Multiple (body)	partial	✗
Video2IMU [3]	Monocular video (2D pose)	Single (waist/hip)	partial	✗
Vi2IMU [8]	Monocular video (2D/3D pose)	Single (wrist)	✗	✗
IMUGPT [4]	Text (LLM) + Pose Sim	Virtual (flexible)	✗	✗
IMUGPT 2.0 [5]	Text (LLM) + Pose Sim (filtered)	Virtual (flexible)	✗	✗
UniMTS [12]	Text + MoCap + LLM	All joints (graph-based)	✗	✗
IMU2Safety (Ours)	Motion Recon + Physics-Based Sim	Wrist-mounted	✓	✓

neural networks to map pose sequences to sensor readings. **Exocentric video** offers key advantages for IMU synthesis: a stable viewpoint capturing the full-body motion and context, easier pose estimation, and a wealth of online data (e.g., Kinetics, UCF) depicting varied movements. Indeed, video-generated IMU data tends to provide more physically accurate motion patterns, whereas text-generated data contributes greater semantic diversity. **Egocentric video** has also been explored recently. For example, COMODO [1] distills knowledge from wearable-camera videos to train an IMU encoder, combining the rich semantics of vision with the efficiency of wrist-mounted sensors. However, *egocentric footage is comparatively scarce and presents challenges like motion blur and partial visibilities due to the moving camera, often making cross-modal mapping less reliable.*

Additionally, current vision-based pipelines also exhibit technical limitations that hinder their applicability to safety-critical IMU synthesis. Most are designed for hip-mounted sensors and do not generalize well to wrist-based configurations common in smartwatches. They often rely solely on 2D pose input, omitting richer motion cues in frames. Furthermore, without explicit physics constraints, generated signals may violate basic inertial consistency—particularly along the vertical axis—and commonly employ architectures like U-Net that are not optimized for modeling long-range temporal dynamics. These limitations motivate our two-stage pipeline, which ensures physical realism, temporal coherence, and wrist-specific fidelity.

Limitations for Rare or Unsafe Behaviors. While recent cross-modal methods have significantly advanced virtual IMU generation, they are primarily designed for routine daily activities and offer limited support for rare or safety-critical behaviors. For example, pose-based approaches such as Video2IMU [3] and Vi2IMU [8] are trained on balanced datasets and typically model high-frequency actions like walking or sitting. Although some datasets include events such as falling or pushing, these are often labeled only by coarse action type, without annotations regarding risk or outcome severity. Text-driven models like IMUGPT [4, 5] and UniMTS [12] expand semantic coverage through language-based conditioning but lack physical grounding and are not explicitly designed to distinguish between safe and unsafe executions of similar actions. COMODO [1] introduces egocentric video-based distillation to improve sensor understanding, yet still focuses on general activity classification without incorporating safety-level annotations. In contrast, our framework explicitly targets safety-critical scenarios and provides ***safety-level labels***, enabling the model to reason not only about the type of action but also its potential risk. This distinction is crucial for applications where understanding the context and severity of human behaviors is essential.

4. Synthetic IMU Signal Generation

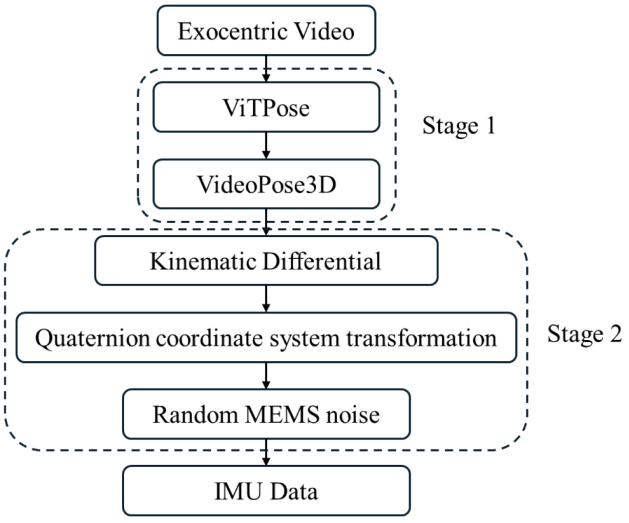

Figure 1 Synthetic IMU Data Generation Pipeline.

The core idea of our technical pipeline is to decompose the complex task of video-to-IMU synthesis into two logically independent and technically mature stages: motion reconstruction and physics-based simulation.

This design enables each stage to leverage the most advanced methods in its respective field. The interface between stages—a 3D skeletal motion sequence—is a standardized data structure with clear physical semantics.

4.1. Stage 1: Motion Reconstruction from Exocentric Video

The goal of this stage is to estimate the 3D coordinates of key human joints (e.g., shoulder, elbow, wrist, hip, knee, ankle) from input 2D video sequences. We will implement a robust and flexible Two-Step Lifting approach.

- **Step 1:** We use the ViTPose 2D pose estimation model. ViTPose and its successors have become benchmarks in pose estimation, demonstrating strong performance in handling occlusion, complex backgrounds, and diverse human tasks. We directly apply this model for inference.
- **Step 2:** We employ VideoPose3D to lift the 2D keypoints to 3D. It utilizes dilated convolutions along the temporal dimension to capture long-range dependencies and resolve the inherent depth ambiguity of single-frame inputs. This results in temporally smoother and physically more plausible 3D pose sequences.

4.2. Stage 2: Physics-Based IMU Simulation for the Wrist

This stage transforms the abstract kinematic data (temporal coordinates and orientation of the wrist joint) from Stage 1 into concrete, physically plausible 6-axis IMU signals (3-axis accelerometer and 3-axis gyroscope). It consists of three modules:

- **Module 1: Kinematic Differentials.** Given time-series 3D wrist positions and forearm orientation, we compute velocity via first-order differentiation and acceleration via second-order differentiation. To mitigate noise amplification from discrete differentiation, we first apply a low-pass filter to the position sequence. Angular velocity is derived from changes in wrist orientation over time.
- **Module 2: Quaternion Coordinate Transformation.** We apply quaternion sandwich multiplication to compute the transformation from the global coordinate system to the sensor's local frame. This step is essential for preserving physical correctness, accurately modeling how a real IMU sensor perceives inertial forces on a moving limb.
- **Module 3: Random MEMS Noise Injection.** While ideal simulations are noise-free, all real MEMS IMUs suffer from inherent stochastic errors. To enhance realism, we inject noise profiles consistent with those observed in physical IMU devices.

5. Data Sources for Wrist-Mounted Devices

To support safety-focused IMU-based HAR, we aim to construct a synthetic dataset covering and seven critical safety-related behaviors from routine daily activities: ①falling, ②losing balance, ③fighting, ④shoving, ⑤dragging, ⑥jumping from height, and ⑦fainting. These actions are synthesized into realistic wrist-mounted and hand-held IMU signals suitable for smartwatch or wristband deployment. The following data sources support different stages of our pipeline (Details please refer to Table 2):

Kinetics-700 Dataset
UCF101 (Amazon Pre-approved)
VIF (Violent Flow)
FineAction
UCF-Crime
XD-Violence
UP-Fall Detection Dataset
Le2i Fall Detection
NTU RGB+D
VIDIMU

6. Stretch Goal

6.1. Model Design Background and Motivation

Current Human Activity Recognition (HAR) research primarily focuses on coarse-grained classification of predefined actions, **lacking the capability for fine-grained behavioral analysis and causal reasoning**. To overcome this limitation, we propose a novel framework called **MHIA** (Multimodal Hierarchical Interaction and Attention), which aims to elevate HAR from simple classification to deeper semantic and causal reasoning.

The core innovations of MHIA include:

- A **Gated Hierarchical Interaction (GHI)** module that enables deep, bidirectional fusion between video and Inertial Measurement Unit (IMU) data streams.
- A **self-supervised multimodal instruction-following data generation pipeline**, which uses a Vision-Language Model (VLM) to automatically produce high-quality instruction–response pairs that align closely with the reasoning goals of the model.

MHIA is an end-to-end trainable framework that utilizes both video and IMU data at inference time, supporting true multimodal understanding. It is expected to perform well on complex reasoning tasks and demonstrate zero-shot generalization to unseen activities.

6.2. Limitations of Existing HAR Research

Existing HAR methods suffer from two key limitations:

- **Modality limitations:** Most models rely on either vision or sensor data alone. Video is susceptible to occlusion, lighting variation, and viewpoint dependency. IMU data captures fine motion but lacks context, object interactions, and semantic clues.
- **Reasoning limitations:** Current models answer the "what" (e.g., walking, running), but lack the ability to infer the "how" and "why"—that is, they cannot reason about execution styles, intentions, or causes.

Although the original motivation stemmed from safety-critical scenarios, this proposal broadens the scope to the

Dataset	Dangerous Actions	IMU Available	Frame-Level Labels	Use for IMU Synthesis	Notes
Kinetics-700	✓(partial)	✗	✗	✓	Large-scale general actions; useful for pose-based IMU generation.
UCF101	✓(partial)	✗	✗	✓	Diverse body movements; good for pose extraction.
VIF (Violent Flow)	✓	✗	✗	✓	Focused on violence detection (e.g., fighting); supports unsafe action modeling.
FineAction	✓	✗	✓	✓	Clear action boundaries; ideal for aligned IMU sequence generation.
UCF-Crime	✓	✗	✗	✓	Real-world surveillance videos; enables dangerous behavior synthesis.
XD-Violence	✓	✗	✓(partial)	✓	Multisource violent/non-violent videos; suited for safety classification.
UP-Fall Detection	✓	✓	✓	✓	Includes synchronized RGB and IMU; useful for training and evaluation.
Le2i Fall Detection	✓	✗	✓	✓	Contains fall events; used for synthetic IMU generation and validation.
NTU RGB+D	✓	✗	✓(pose)	✓	High-quality 3D pose; ideal for physically consistent IMU synthesis.

Table 2 Summary of video datasets considered for synthetic IMU generation.

more general and challenging problem of **Fine-Grained Activity Reasoning**.

6.3. MHIA Framework Overview

MHIA is an end-to-end system where raw video (RGB frames) and 6-axis IMU sequences are first encoded by modality-specific encoders. The resulting features are fed into the core **GHI module** for deep fusion. A projection layer then maps the fused representation into the embedding space of a Large Language Model (LLM), which performs inference and generation (e.g., description, QA, reasoning).

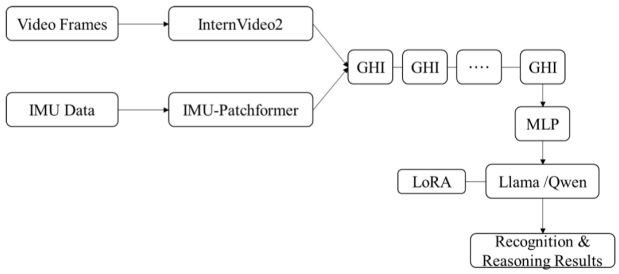

Figure 2 Recognition Model Architecture.

6.4. Spatiotemporal Video Encoder

We use a pretrained InternVideo2-6B backbone. Videos are sparsely sampled (e.g., 2 FPS for 16 frames). Multi-layer features from InternVideo's ViT backbone capture a hierarchy from low-level motion to high-level semantics.

6.4.1. IMU Kinematics Encoder (IMU-Patchformer)

This encoder is based on PatchTST and processes raw 6-axis IMU data (3-axis accelerometer + 3-axis gyroscope). The signal is segmented into non-overlapping patches (e.g., 32 time steps with stride 16).

We introduce a **channel-wise attention module** before the transformer to model dependencies across the six channels. At each time step, attention is applied across all channels (e.g., X/Y/Z acceleration and angular velocity). The output is then passed to a transformer to model temporal dynamics.

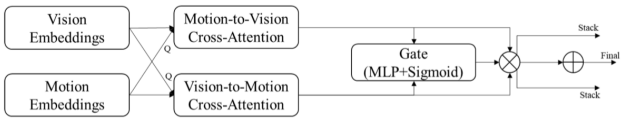

Figure 3 GHI Model Architecture.

6.4.2. Gated Hierarchical Interaction (GHI) Module

Each GHI block performs bidirectional cross-modal fusion:

- **Motion-to-Vision Cross-Attention:** IMU patch embeddings are queries; video feature maps are keys/values. This grounds motion spikes in visual evidence.
- **Vision-to-Motion Cross-Attention:** Global pooled video features are queries; IMU sequences are keys/values. This gives global context to sensor data.

A small MLP-based **gating network** produces a dynamic weight $g \in [0, 1]$, which controls the weighted sum of the two refined features. This allows the model to adapt to noisy or unreliable modality inputs and solves the problem

of modality imbalance.

6.4.3. LLM Integration and Instruction Following

The fused representation is linearly projected into the token embedding space of an LLM (e.g., LLaMA-3-8B). We train the full system end-to-end, and fine-tune the LLM efficiently using **LoRA** (Low-Rank Adaptation).

6.4.4. Self-Supervised Instruction Data Generation Pipeline

Standard HAR datasets (e.g., UTD-MHAD, SisFall) provide only class labels (e.g., "walking"), which are insufficient for complex reasoning.

We propose a pipeline that uses a closed-source VLM (e.g., GPT-4V or Gemini) as the **teacher model** to generate instruction–response pairs.

Inputs to the teacher:

(a) Synchronized video clips
(b) IMU time-series plots
(c) Ground truth action label

Three prompt types:

- **Fine-grained Description**
  Instruction: "Provide a detailed technical description."
  Response: "The subject performs a lunge. Right leg steps forward, producing a positive Y-axis acceleration peak and negative Z-axis gyroscope spike."
- **Causal and Counterfactual Reasoning**
  Instruction: "Explain the cause of the fall and compare it to a safe motion."
  Response: "The fall was likely caused by tripping, indicated by sudden uncontrolled acceleration and high angular velocity. In contrast, a controlled sit-to-stand movement shows smoother vertical acceleration and minimal rotation."
- **Comparative Analysis**
  Instruction: "Compare the two walking gaits."
  Response: "The first is steady flat walking with low-amplitude oscillation; the second is uphill, showing higher amplitude and lower cadence."

References

[1] Baiyu Chen, Wilson Wongso, Zechen Li, Yonchanok Khaokaew, Hao Xue, and Flora Salim. Comodo: Cross-modal video-to-imu distillation for efficient egocentric human activity recognition. *arXiv preprint arXiv:2503.07259*, 2025.

[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*, 4(3):1–29, 2020.

[3] Arttu Lämsä, Jaakko Tervonen, Jussi Liikka, Constantino Álvarez Casado, and Miguel Bordallo López. Video2imu: Realistic imu features and signals from videos. In *2022 IEEE-EMBS International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pages 1–5. IEEE, 2022. Available: https://arxiv.org/pdf/2202.06547.

[4] Zikang Leng, Hyeokhyen Kwon, and Thomas Plötz. Generating virtual on-body accelerometer data from virtual textual descriptions for human activity recognition. In *Proceedings of the 2023 ACM International Symposium on Wearable Computers*, pages 39–43, 2023.

[5] Zikang Leng, Amitrajit Bhattacharjee, Hrudhai Rajasekhar, Lizhe Zhang, Elizabeth Bruda, Hyeokhyen Kwon, and Thomas Plötz. Imugpt 2.0: Language-based cross modality transfer for sensor-based human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2024. Available: https://arxiv.org/pdf/2402.01049.

[6] Hiroki Ohashi, M Al-Nasser, Sheraz Ahmed, Takayuki Akiyama, Takuto Sato, Phong Nguyen, Katsuyuki Nakamura, and Andreas Dengel. Augmenting wearable sensor data with physical constraint for dnn-based human-action recognition. In *ICML 2017 times series workshop*, pages 6–11, 2017.

[7] Vitor Fortes Rey, Peter Hevesi, Onorina Kovalenko, and Paul Lukowicz. Let there be imu data: generating training data for wearable, motion sensor based activity recognition from monocular rgb videos. In *Adjunct proceedings of the 2019 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2019 ACM international symposium on wearable computers*, pages 699–708, 2019.

[8] Panneer Selvam Santhalingam, Parth Pathak, Huzefa Rangwala, and Jana Kosecka. Synthetic smartwatch imu data generation from in-the-wild asl videos. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(2):1–34, 2023.

[9] Odongo Steven Eyobu and Dong Seog Han. Feature representation and data augmentation for human activity classification based on wearable imu sensor data using a deep lstm neural network. *Sensors*, 18(9):2892, 2018.

[10] Nilay Tufek, Murat Yalcin, Mucahit Altintas, Fatma Kalaoglu, Yi Li, and Senem Kursun Bahadir. Human action recognition using deep learning methods on limited sensory data. *IEEE Sensors Journal*, 20(6):3101–3112, 2019.

[11] Qingsong Wen, Liang Sun, Fan Yang, Xiaomin Song, Jingkun Gao, Xue Wang, and Huan Xu. Time series data augmentation for deep learning: A survey. *arXiv preprint arXiv:2002.12478*, 2020.

[12] Xiyuan Zhang, Diyan Teng, Ranak Roy Chowdhury, Shuheng Li, Dezhi Hong, Rajesh Gupta, and Jingbo Shang. Unimts: Unified pre-training for motion time series. *Advances in Neural Information Processing Systems*, 37:107469–107493, 2024.