

1 **PrimeIMU: High-Frequency Video-to-IMU Synthesis via Physics-Guided**
2 **Simulation and Hybrid U-Net Refinement**
3

4 ANONYMOUS AUTHOR(S)
5

6 Inertial Measurement Units (IMUs) are fundamental sensing components across robotics, drones, and other embodied systems,
7 providing high-frequency inertial signals essential for control, stabilization, state estimation, and dynamics modeling. However,
8 acquiring large-scale, high-sampling-rate IMU datasets for diverse articulated platforms remains challenging: deploying, calibrating,
9 and synchronizing physical IMU sensors is costly, time-consuming, and often constrained by hardware limitations and operational
10 requirements. To alleviate these challenges, recent work has explored synthesizing IMU signals from either model-synthesized motions
11 or vision-derived kinematic estimates. While text-to-motion-based IMU synthesis approaches scale well, they lack real-world visual
12 grounding. Vision-driven video-to-IMU approaches offer such grounding but are fundamentally constrained by the limited temporal
13 resolution of visual inputs and the inherent mismatch between kinematic estimates and inertial measurements.
14

15 Building on these observations, we propose PrimeIMU, a framework that synthesizes IMU signals from vision-derived kinematic
16 trajectories and physics-simulated IMU signals, using a hybrid U-Net that learns to map simulated signals to real sensor measurements.
17 Extensive experiments demonstrate that PrimeIMU (1) improves the quality of inertial signal synthesis, (2) generalizes to unseen
18 motion patterns while preserving realistic signal characteristics, (3) enables synthetic-only training of downstream models with
19 performance close to real-sensor training and (4) yields further gains when used to augment real data, and (5) adapts across different
20 datasets and sensor configurations with minimal fine-tuning, supporting downstream tasks under domain shift.
21

22 CCS Concepts: • Computing methodologies → Machine learning.
23

24 Additional Key Words and Phrases: IMU Signal Synthesis
25

26 **ACM Reference Format:**
27

28 Anonymous Author(s). 2018. PrimeIMU: High-Frequency Video-to-IMU Synthesis via Physics-Guided Simulation and Hybrid U-Net
29 Refinement. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym*
30 'XX). ACM, New York, NY, USA, 24 pages. <https://doi.org/XXXXXXX.XXXXXXX>

31 **1 Introduction**
32

33 Inertial Measurement Unit (IMU) data are widely used across articulated mechanical and robotic systems [18, 19, 25, 34],
34 including aerial drones [7, 26] and other embodied platforms [1, 8]. High-frequency inertial measurements capture
35 fine-grained motion dynamics [19, 34], supporting tasks such as motion analysis, dynamic behavior characterization [25],
36 operational assessment [38], and model-based simulation of articulated structures [24]. Achieving these applications at
37 scale requires large, high-frequency IMU datasets with reliable annotations. However, collecting such data remains highly
38 challenging, as the process is time-consuming, costly, and requires substantial domain expertise [5, 11, 15, 23, 27, 36, 39].
39

40 To address the scarcity of labeled IMU data, prior work has explored synthetic IMU generation for data augmentation
41 through two primary paradigms. The first paradigm leverages *large language models* (LLMs) to generate diverse motion
42 descriptions, which are converted into 3D articulated trajectories via motion synthesis and subsequently transformed
43

44 Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not
45 made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components
46 of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on
47 servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
48

49 © 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
50

51 Manuscript submitted to ACM
52

53 Manuscript submitted to ACM

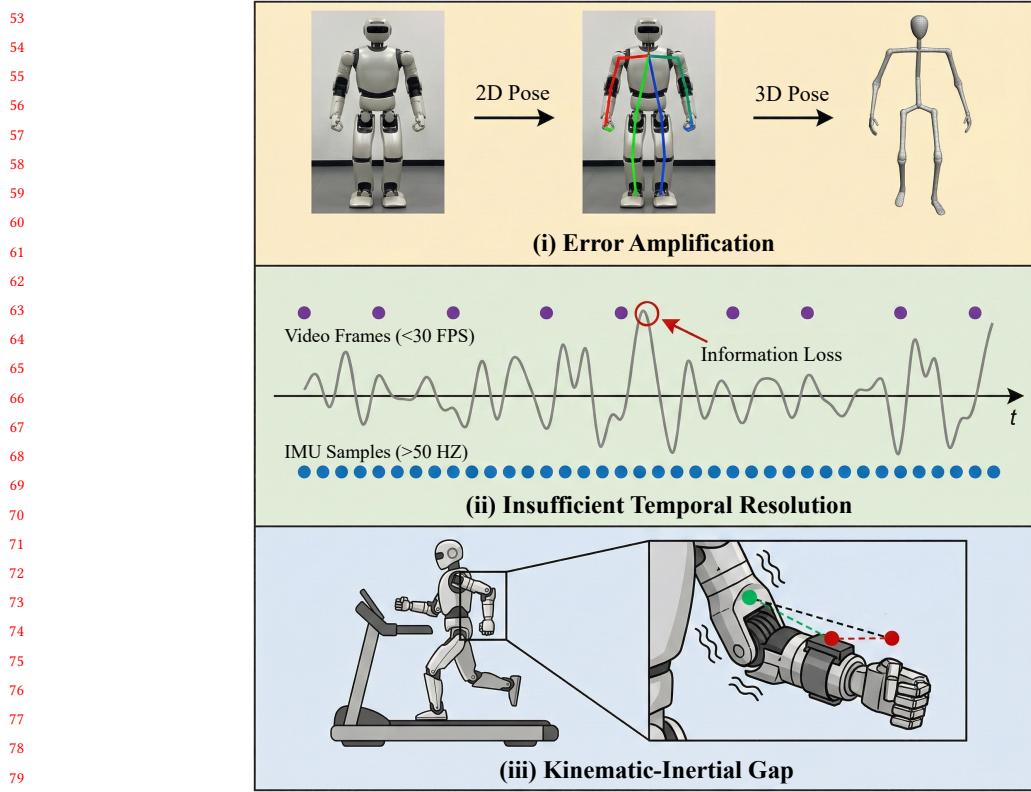


Fig. 1. Illustration of three fundamental limitations in current vision-driven IMU synthesis pipelines: (i) error amplification introduced by kinematic estimation, (ii) insufficient temporal resolution when projecting low-frame-rate video into high-frequency IMU signals, and (iii) discrepancies between idealized kinematic models and the motion sensed by an IMU mounted on a moving structure.

into virtual IMU signals through motion-to-IMU modeling [15, 16, 41]. The second paradigm relies on *vision-driven models* to extract 2D kinematic cues from visual inputs, lift them to 3D articulated representations, and synthesize virtual IMU signals via kinematic modeling [13, 14, 29]. While LLM-based approaches offer strong scalability, they lack visual grounding and often produce IMU signals with unrealistic dynamics. Vision-driven methods provide visual grounding, but still fall short of real IMU quality for three fundamental reasons (Figure 1). *First*, errors in visual kinematic estimation propagate and amplify through temporal differentiation, resulting in noisy or spiky inertial signals. *Second*, visual frame rates (typically 15–30 FPS) are substantially lower than IMU sampling rates ($> 50\text{Hz}$), leading to irreversible temporal under-resolution that cannot be recovered through interpolation. *Third*, a persistent kinematic–inertial gap exists between idealized articulated motion and the inertial response of physical IMU hardware, whose measurements reflect mounting dynamics and other sensor-specific non-idealities. Together, these limitations substantially reduce the realism of synthesized IMU signals and limit the effectiveness of synthetic data for downstream learning tasks.

Addressing these challenges would provide significant value for both research and real-world deployment, enabling the development of IMU-driven models with minimal reliance on large-scale real sensor data. Consequently, one crucial question has arisen: *Is it feasible to synthesize IMU signals that faithfully capture the physical behavior of real sensors while remaining effective for IMU-based downstream applications?*

IMU Generation Method	Realism-Oriented Designs				
	real.	vis.+phys.	temp.res.	dev.adpt.	kin.t.freq.
IMUGPT [16]	X	X	X	X	X
IMUGPT 2.0 [15]	X	X	X	X	X
UniMTS [41]	X	X	X	X	X
Video2IMU [14]	X	X	X	X	X
Vi2IMU [29]	X	✓	X	X	X
IMUTube [13]	X	✓	X	X	X
PrimeIMU	✓	✓	✓	✓	✓

Table 1. Comparison of PrimeIMU with State-of-the-Art IMU generation methods on realism-oriented designs, where *real.*: explicitly aims for realistic IMU generation (beyond pure augmentation); *vis.+phys.*: combines visual grounding with physical modeling; *temp.res.*: addresses limited temporal resolution from video frame rates; *dev.adpt.*: adapts synthesis to device-specific characteristics (e.g., FPS, IMU sampling rates); *kin.t.freq.*: incorporates kinematic, temporal, and frequency-domain information during synthesis.

To address this question, we propose PrimeIMU, a framework for synthesizing realistic IMU signals from video-derived kinematics that explicitly targets key realism-oriented design challenges, including physical consistency, temporal resolution mismatch, and device-specific sensor behavior (Table 1). PrimeIMU first performs physics-based IMU simulation on video-derived kinematic trajectories and then refines the simulated signals using a hybrid U-Net to bridge the gap between idealized kinematics and real sensor measurements. Our contributions are as follows:

- **Realistic, Sensor-Faithful Video-to-IMU Synthesis.** We propose PrimeIMU, a video-driven IMU synthesis pipeline for generating realistic and sensor-faithful inertial signals. PrimeIMU integrates physics-based IMU simulation with a hybrid U-Net refinement module, jointly addressing three fundamental limitations of prior approaches: noise amplification from pose-derived kinematics, temporal under-resolution caused by low video frame rates, and the mismatch between idealized kinematic modeling and the behavior of physical IMU sensors.
- **Generalization to Unseen Motion Patterns.** We demonstrate that PrimeIMU generalizes to motion patterns not observed during training, suggesting that it captures underlying inertial structure beyond the training distribution.
- **Synthetic-Only Training of IMU-Based Models.** We show that downstream models trained solely on PrimeIMU-generated IMU signals achieve performance comparable to models trained on real sensor data, indicating that high-fidelity synthetic IMUs can substitute for real measurements when sensor data are limited.
- **Augmentation Benefits with Real IMU Data.** PrimeIMU-generated signals introduce sensor-faithful inertial patterns that complement real sensor measurements by expanding physically plausible inertial variations, leading to consistent performance gains in downstream tasks when used for data augmentation.
- **Cross-Dataset and Cross-Device Adaptation.** PrimeIMU adapts effectively to domain shifts across datasets and IMU hardware, requiring minimal fine-tuning to adapt to new sensor characteristics and sampling configurations.

2 Related Work

2.1 Data Scarcity and Learning Paradigms

Despite the widespread deployment of IMUs in articulated mechanical and robotic systems, the availability of large-scale, diverse, and well-calibrated inertial datasets remains limited. Constructing datasets that span a broad range of motion patterns, operating conditions, and device configurations requires controlled hardware deployment, precise

calibration, and repeated data collection. These requirements impose substantial operational overhead and often lead to datasets that are narrow in scope and insufficiently diverse for robust inertial modeling. Moreover, variations in device characteristics and sensor mounting conditions introduce nontrivial device-dependent distribution shifts, further hindering cross-system generalization [10, 17]. As a result, existing IMU datasets remain constrained in operational coverage despite the ubiquity of micro-electro-mechanical systems (MEMS)-based sensors.

Beyond data quantity, annotation quality presents an additional bottleneck. Establishing reliable ground truth for IMU data typically relies on structured collection protocols or manual post-processing of continuous recordings, which can introduce inconsistencies in activity boundaries, temporal alignment, and labeling conventions [5, 11, 15, 23, 27, 36, 39]. Such variability further limits the scalability and uniformity of supervised IMU training corpora.

To mitigate these challenges, prior work has explored alternative learning paradigms, including self-supervised pretraining [12, 28], few-shot learning and domain adaptation [3, 9], metric-based representations [6], adversarial robustness techniques [2], and cross-dataset transfer [30]. While these approaches improve generalization under data scarcity, they remain fundamentally constrained by the diversity and physical coverage of real IMU measurements.

This limitation motivates the development of realistic synthetic IMU generation, where synthesized signals explicitly aim to reproduce device-level inertial behavior while scaling across motion types and hardware configurations. Such synthetic data can complement limited real measurements and enable effective downstream inertial modeling when large-scale physical data collection is impractical.

2.2 Cross-Modality IMU Synthesis

Owing to the difficulty of collecting large-scale annotated IMU datasets, recent work has increasingly explored the synthesis of inertial signals from alternative sensing modalities. Existing approaches broadly fall into two categories: text-driven and video-driven synthesis.

Text-driven pipelines, such as IMUGPT [16] and IMUGPT 2.0 [15], leverage large language models to generate diverse motion descriptions, which are then mapped to parametric trajectories and converted into synthetic IMU sequences. UniMTS [41] further incorporates linguistic priors before translating text descriptions into skeletal motions. While these approaches offer strong scalability and diversity, they lack visual grounding and often struggle to preserve the statistical and physical characteristics of real IMU measurements.

Video-driven pipelines introduce visual grounding by extracting motion cues directly from videos. IMUTube [13] estimates 2D poses, lifts them to 3D, and differentiates the resulting trajectories to approximate inertial signals. Video2IMU [14] directly regresses IMU sequences from 2D pose representations, while Vi2IMU [29] focuses on wrist-mounted sensors in constrained communication scenarios. Compared to text-driven methods, video-driven approaches generally yield more realistic inertial signatures due to their explicit visual grounding.

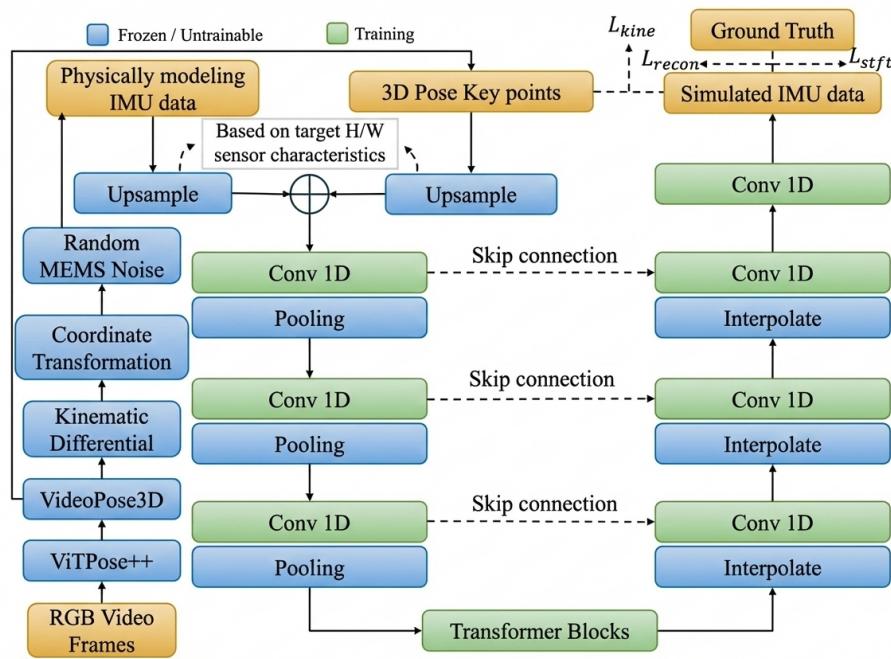
Despite these advantages, video-driven IMU synthesis remains fundamentally limited. Errors in 2D and 3D pose estimation are amplified through temporal differentiation, resulting in unstable or noisy inertial predictions. In addition, the inherent mismatch between video frame rates and IMU sampling frequencies constrains the synthesis of high-frequency motion components. More importantly, visually derived kinematic trajectories alone cannot fully capture the inertial dynamics measured by physical IMU sensors, whose outputs depend on forces, vibrations, and mounting-dependent effects that are not represented in purely kinematic models. Together, these limitations indicate that effective IMU synthesis requires jointly leveraging visual kinematics and physics-based inertial modeling to more faithfully approximate the behavior of real IMU sensors.

209 **3 Method**

210 Our objective is to synthesize high-frequency and physically plausible inertial measurement unit (IMU) signals from
 211 standard RGB video inputs. This problem is challenging due to the substantial domain gap between low-frame-rate
 212 visual observations and high-frequency inertial measurements, as well as the noise and estimation errors inherent in
 213 vision-based motion capture.
 214

215 To address these challenges, we propose PrimeIMU, a two-stage synthesis framework illustrated in Figure 2. In the
 216 first stage, visual motion is estimated from RGB videos and translated into an initial IMU signal using physics-based
 217 kinematic modeling. This stage captures coarse motion trends but remains limited by low temporal resolution and
 218 accumulated estimation noise. In the second stage, a hybrid U-Net-based generative refinement module reconstructs
 219 high-frequency inertial dynamics, suppresses noise, and enforces kinematic consistency at the target IMU sampling
 220 rate, producing realistic and sensor-faithful IMU signals.
 221

222
 223
 224
 225



250 Fig. 2. Overview of the PrimeIMU pipeline for synthesizing high-frequency IMU signals from RGB videos. **(Stage 1: Physics-based**
 251 **Simulation)** Each RGB frame is processed by ViTPose++ to estimate 2D keypoints, which are then lifted to 3D joint trajectories using
 252 VideoPose3D. The 3D poses are converted into low-frequency simulated IMU signals through kinematic differentiation, coordinate
 253 transformation, and MEMS noise modeling. **(Stage 2: Deep Generative Refinement)** A Hybrid U-Net with a transformer bottleneck
 254 takes the upsampled simulated IMU together with the upsampled pose sequence, refines them via convolutional-transformer blocks
 255 with skip connections, and generates realistic high-frequency IMU signals. Training is guided by a composite loss that combines
 256 time-domain reconstruction, frequency-domain STFT alignment, and kinematic consistency to enforce both fidelity and physical
 257 plausibility.

258
 259
 260

261 3.1 Stage 1: Physics-Guided IMU Simulation

262 The first stage of PrimeIMU converts the input video into a coarse IMU estimate by extracting 3D kinematic trajectories
 263 and applying a physics-guided inertial simulation model. This stage provides a physically grounded approximation of
 264 the target sensor signals that captures coarse motion trends, but remains low-frequency and noisy due to limitations in
 265 visual estimation accuracy and video frame rates.
 266

267 *3.1.1 3D Pose Estimation.* Given an RGB video sequence, our first step is to reconstruct the 3D trajectory of key object
 268 joints. We adopt a robust two-step lifting approach. First, we process each video frame with ViTPose++ [40], a powerful
 269 vision transformer-based model, to obtain accurate 2D keypoint coordinates for each object in the scene. Subsequently,
 270 the resulting 2D pose sequence is fed into VideoPose3D [22], a temporal convolutional network that leverages motion
 271 context to lift the 2D coordinates into a coherent 3D skeleton sequence. The output of this step is a sequence of 3D
 272 joint positions $P_{3D} \in \mathbb{R}^{T_v \times K \times 3}$, where T_v is the number of video frames and K is the number of keypoints.
 273

274 *3.1.2 Physics-Driven Kinematic Modeling.* With the reconstructed 3D joint trajectories, we simulate the readings of a
 275 6-axis IMU (3-axis accelerometer and 3-axis gyroscope) attached to a specific object segment. This simulation consists
 276 of three components: kinematic differentiation, coordinate system transformation, and MEMS noise modeling.
 277

278 *Kinematic Differentiation.* This module computes the linear acceleration and angular velocity from the pose data. Let
 279 the 3D position of the distal joint of a segment (e.g., the wrist) at time t be denoted by the vector $\mathbf{p}(t)$. The global linear
 280 velocity $\mathbf{v}(t)$ and acceleration $\mathbf{a}_{\text{global}}(t)$ are obtained by taking the first and second time derivatives of the position,
 281 respectively:
 282

$$283 \quad \mathbf{v}(t) = \frac{d\mathbf{p}(t)}{dt}, \quad \mathbf{a}_{\text{global}}(t) = \frac{d^2\mathbf{p}(t)}{dt^2}. \quad (1)$$

284 Directly differentiating discrete, noisy position data drastically amplifies noise. Therefore, we first apply a Savitzky-
 285 Golay filter [31] to the position sequence $\mathbf{p}(t)$ to obtain a smoothed trajectory before differentiation, which is imple-
 286 mented using finite differences.
 287

288 The orientation of the object segment is more robustly defined by the vector connecting the proximal joint $\mathbf{p}_{\text{prox}}(t)$
 289 to the distal joint $\mathbf{p}_{\text{dist}}(t)$. We use this vector to construct a time-varying local coordinate frame for the segment,
 290 represented by a sequence of orientation quaternions $q(t)$. The local angular velocity $\boldsymbol{\omega}_{\text{local}}(t)$ is then derived from
 291 the temporal evolution of these quaternions. Given the relationship $\dot{q}(t) = \frac{1}{2} \begin{bmatrix} 0 \\ \boldsymbol{\omega}_{\text{local}}(t) \end{bmatrix} \otimes q(t)$, where \otimes denotes
 292 quaternion multiplication, the angular velocity can be computed as:
 293

$$294 \quad \begin{bmatrix} 0 \\ \boldsymbol{\omega}_{\text{local}}(t) \end{bmatrix} = 2 \cdot \dot{q}(t) \otimes q(t)^{-1}, \quad (2)$$

295 where $q(t)^{-1}$ is the quaternion conjugate and $\dot{q}(t)$ is computed via finite differences.
 296

297 *Coordinate System Transformation.* An IMU measures acceleration and angular velocity in its own local coordinate
 298 frame. The angular velocity $\boldsymbol{\omega}_{\text{local}}(t)$ is already in this frame. However, the linear acceleration $\mathbf{a}_{\text{global}}(t)$ must be
 299 transformed. Using the orientation quaternion $q(t)$, we rotate the global acceleration vector into the local frame and
 300 subtract the effect of gravity, $\mathbf{g} = [0, 0, -9.81]^T \text{m/s}^2$, which is also transformed into the local frame:
 301

$$302 \quad \mathbf{a}_{\text{local}}(t) = R(q(t))^{-1}(\mathbf{a}_{\text{global}}(t)) - R(q(t))^{-1}\mathbf{g}. \quad (3)$$

313 where $R(q(t))$ is the rotation matrix corresponding to quaternion $q(t)$. This step is crucial for simulating realistic
 314 accelerometer readings that include gravitational components.
 315

316 *MEMS Noise Modeling.* Real IMU sensors are subject to various sources of noise. To enhance the realism of our
 317 simulated data, we model two primary noise types: a Gaussian white noise component and a random walk bias. The
 318 final simulated IMU signal $I_{\text{sim}}(t) = [\mathbf{a}_{\text{sim}}(t), \boldsymbol{\omega}_{\text{sim}}(t)]$ is generated as:
 319

$$\mathbf{a}_{\text{sim}}(t) = \mathbf{a}_{\text{local}}(t) + \mathbf{n}_a(t) + \mathbf{b}_a(t), \quad (4)$$

$$\boldsymbol{\omega}_{\text{sim}}(t) = \boldsymbol{\omega}_{\text{local}}(t) + \mathbf{n}_{\omega}(t) + \mathbf{b}_{\omega}(t), \quad (5)$$

320 where $\mathbf{n}(t)$ is sampled from a zero-mean Gaussian distribution whose standard deviation depends on the sensor's noise
 321 density, and $\mathbf{b}(t)$ is a bias term that evolves as a random walk. The output of this stage is a low-frequency simulated
 322 IMU signal, $I_{\text{sim}} \in \mathbb{R}^{T_v \times 6}$.
 323

324 3.2 Stage 2: Hybrid U-Net Generative Refinement

325 The physics-guided simulation in Stage 1 produces an initial IMU estimate I_{sim} that is physically grounded but remains
 326 noisy and limited to the low temporal resolution of the source video. To bridge the gap between this coarse approximation
 327 and realistic high-frequency inertial measurements, the second stage employs a generative refinement network that
 328 learns to map low-quality simulated signals to high-quality IMU data, conditioned on the corresponding low-frequency
 329 pose sequence.
 330

331 *3.2.1 Network Architecture.* We design a Hybrid U-Net architecture that combines the local feature extraction power of
 332 convolutions with the global context modeling of transformers. The network takes a low-frequency pose sequence
 333 $P_{\text{low}} \in \mathbb{R}^{T_l \times D_{\text{pose}}}$ and a random noise vector $N \in \mathbb{R}^{T_h \times D_{\text{imu}}}$ as input, where T_l and T_h are the lengths of the low- and
 334 high-frequency sequences, respectively.
 335

336 The pose sequence P_{low} is first upsampled to the target length T_h via linear interpolation and then projected into a
 337 high-dimensional embedding space. This pose embedding is added to an embedding of the input noise N . The resulting
 338 sequence $X \in \mathbb{R}^{T_h \times D_{\text{model}}}$ is the input to the U-Net.
 339

340 *Encoder.* The encoder consists of a sequence of 1D convolutional blocks with residual connections, each followed by
 341 max-pooling to progressively reduce temporal resolution. The feature maps produced before pooling are stored as skip
 342 connections, preserving fine-grained temporal details for later reconstruction.
 343

344 *Bottleneck.* At the lowest temporal resolution, the encoded features are passed through a Transformer block. This
 345 bottleneck enables the model to capture long-range temporal dependencies and global motion patterns that are difficult
 346 to model using purely convolutional operations. Positional encodings are added prior to the Transformer to retain
 347 temporal ordering information.
 348

349 *Decoder.* The decoder mirrors the encoder's structure. At each stage, the feature map is upsampled using linear
 350 interpolation. The upsampled features are concatenated with the corresponding skip connection from the encoder
 351 path and passed through a 1D convolutional block. This process progressively refines the features while restoring the
 352 original temporal resolution. A final 1D convolutional layer projects the output features back to the dimension of the
 353 IMU data, yielding the generated high-frequency signal $I_{\text{gen}} \in \mathbb{R}^{T_h \times 6}$.
 354

365 3.3 Composite Loss Design

366
367 The generative refinement network is trained end-to-end using a composite loss that jointly evaluates signal fidelity
368 in the time and frequency domains while enforcing physical consistency with the input motion. The ground-truth
369 supervision is provided by real high-frequency IMU measurements $I_{\text{gt}} \in \mathbb{R}^{T_h \times 6}$.
370

371 *Reconstruction Loss ($\mathcal{L}_{\text{recon}}$)*. To enforce time-domain agreement between the generated signal and the ground truth,
372 we apply an L1 reconstruction loss:
373

$$\mathcal{L}_{\text{recon}} = \|I_{\text{gen}} - I_{\text{gt}}\|_1. \quad (6)$$

374 This term penalizes deviations in signal amplitude and temporal structure.
375

376
377 *STFT Loss ($\mathcal{L}_{\text{stft}}$)*. Accurately reproducing the frequency content of IMU signals is critical for realism. We therefore
378 incorporate a multi-resolution Short-Time Fourier Transform (STFT) loss that aligns the spectral characteristics of the
379 generated and real signals. The loss consists of a spectral convergence term and a log-magnitude term, computed across
380 multiple Fast Fourier Transform (FFT) resolutions:
381

$$\mathcal{L}_{\text{stft}} = \sum_{r \in R} \left(\frac{\| |S_r(I_{\text{gen}})| - |S_r(I_{\text{gt}})| \|_F}{\| |S_r(I_{\text{gt}})| \|_F} + \|\log |S_r(I_{\text{gen}})| - \log |S_r(I_{\text{gt}})| \|_1 \right), \quad (7)$$

382 where $S_r(\cdot)$ denotes the STFT operator at resolution r , $|\cdot|$ is the magnitude, and $\|\cdot\|_F$ is the Frobenius norm.
383

384
385 *Kinematic Consistency Loss ($\mathcal{L}_{\text{kine}}$)*. To ensure that the generated IMU signals remain physically consistent with the
386 input motion, we introduce a kinematic consistency loss. Specifically, we extract the acceleration component of the
387 generated signal, $i_{\text{gen}}^{\text{accel}} \in \mathbb{R}^{T_h \times 3}$, and apply a differentiable double integration operation to recover a high-frequency 3D
388 trajectory \hat{P}_{high} . An initial position from the ground-truth low-frequency pose is used as the integration constant. The
389 recovered trajectory is then downsampled to the original low temporal resolution, producing \hat{P}_{low} . The kinematic loss
390 is defined as the Mean Squared Error (MSE) between this trajectory and the input low-frequency pose P_{low} that was an
391 input to the network:
392

$$\mathcal{L}_{\text{kine}} = \|\text{Downsample}(\hat{P}_{\text{high}}) - P_{\text{low}}\|_2^2. \quad (8)$$

393 This term regularizes the generative process by enforcing consistency with the underlying kinematic motion.
394

395 *Total Loss. ($\mathcal{L}_{\text{total}}$)*. The overall training objective is a weighted sum of the three loss terms:
396

$$\mathcal{L}_{\text{total}} = \lambda_{\text{recon}} \mathcal{L}_{\text{recon}} + \lambda_{\text{stft}} \mathcal{L}_{\text{stft}} + \lambda_{\text{kine}} \mathcal{L}_{\text{kine}}, \quad (9)$$

397 where λ_{recon} , λ_{stft} , and λ_{kine} are hyper-parameters that balance the contribution of each term.
398

400 4 Experiments

401 4.1 Setup

402 *4.1.1 Datasets*. We conduct experiments on two publicly available multimodal benchmarks that provide both video
403 and inertial sensor modalities for activity recognition: *(i) UTD-MHAD* [4] contains synchronized RGB videos, depth
404 videos, skeleton joint positions, and 6-axis inertial signals (accelerometer and gyroscope) recorded from a Kinect
405 camera and a single wearable inertial sensor. Eight subjects performed 27 different actions, including sports (e.g.,
406 basketball shoot, tennis swing), hand gestures (e.g., wave, clap, draw shapes), daily activities (e.g., sit-to-stand, knock
407 on door), and exercises (e.g., squat, lunge, arm curl). Each subject repeated every action four times, yielding a total of
408 409
410 411
412 413
413
414 415
415
416 Manuscript submitted to ACM

417 861 sequences. The inertial sensor was placed on the right wrist or thigh depending on the action, and all modalities
 418 were temporally synchronized and manually segmented, making UTD-MHAD a compact yet diverse benchmark for
 419 multimodal activity recognition. (ii) **MM-Fit** [33] contains synchronized multi-view RGB-D videos, skeleton poses, and
 420 wearable inertial signals (accelerometer, gyroscope, and magnetometer). Ten subjects performed ten types of resistance
 421 exercises, including squats, lunges, push-ups, sit-ups, curls, rows, presses, raises, and jumping jacks. Each exercise
 422 consisted of three sets of ten repetitions, resulting in 21 sessions, 616 sets, and 6160 annotated repetitions. All modalities
 423 were time-synchronized using a calibration jump and annotated with exercise types, set boundaries, and repetition
 424 counts, providing a challenging and diverse benchmark for multimodal activity recognition in workout scenarios.
 425

426 4.1.2 *Baseline*. Because no prior state-of-the-art video-to-IMU methods publicly release their code, we construct a
 427 reproducible baseline by following the methodological descriptions provided in recent work, in particular, IMUTube
 428 [13] and UniMTS [41]. IMUTube is specifically designed for video-to-IMU synthesis, and we therefore re-implement its
 429 pipeline as faithfully as the available details permit. Although UniMTS is not a video-to-IMU approach, several of its
 430 design components for multimodal motion-to-sensor generation inform aspects of our IMUTube re-implementation.
 431 Unless otherwise stated, all references to the *Baseline* refer to this reconstructed IMUTube pipeline.
 432

433 4.1.3 *IMU Generation Quality Evaluation Metrics*. We assess the fidelity of generated IMU signals by comparing them to
 434 ground-truth recordings using both error- and correlation-based measures. Error-based metrics (RMSE/MAE) quantify
 435 magnitude discrepancies but may overlook trend misalignment, while correlation-based metrics (R^2 and Pearson)
 436 ensure temporal and physical plausibility: (i) **Root Mean Square Error (RMSE)** emphasizes large deviations between
 437 generated and real signals, making it sensitive to spikes; small values indicate stable predictions, while large values
 438 suggest spiky or unstable outputs. (ii) **Mean Absolute Error (MAE)** captures the average absolute deviation from
 439 ground truth; lower values reflect better overall alignment, while higher values indicate consistent drift. (iii) **Coefficient**
 440 **of Determination (R^2)** measures the proportion of variance in the real signal explained by the generated one; values
 441 close to 1 indicate faithful reproduction of dynamics, values near 0 imply mean-level prediction, and negative values
 442 denote worse-than-mean performance. (iv) **Pearson Correlation** quantifies linear trend alignment between generated
 443 and ground-truth signals; values near +1 denote strong positive alignment, values near 0 indicate no relation, and
 444 values near -1 imply reversed, physically implausible trends.
 445

446 4.1.4 *Classification Performance Evaluation Metrics*. We adopt two standard metrics for action recognition evaluation:
 447 (i) **Accuracy (Acc)** is the overall Top-1 classification accuracy, i.e., the fraction of correctly predicted activity labels
 448 across all samples. (ii) **Macro F1 (F1)** is the unweighted average of per-class F1 scores, giving equal importance to each
 449 class and thus providing robustness to class imbalance.
 450

451 4.1.5 *Classification Models*. We evaluate the effectiveness of synthetic IMU data on five representative models spanning
 452 both classical machine learning and deep learning paradigms: (i) **Random Forest** [21] is an ensemble of decision trees
 453 that serves as a strong non-deep-learning baseline for IMU classification. (ii) **Support Vector Machine (SVM)** [35] is a
 454 kernel-based margin classifier that provides a competitive traditional baseline for time-series data. (iii) **DeepConvLSTM**
 455 [20] integrates convolutional layers for local feature extraction with LSTMs for temporal sequence modeling and
 456 is widely used in action recognition tasks. (iv) **DeepConvLSTM_Attention** [32] extends DeepConvLSTM by adding a
 457 self-attention mechanism to capture long-range temporal dependencies through adaptive weighting of time steps. (v)
 458 **Transformer** [37] is a fully attention-based architecture that models long-range dependencies without recurrence
 459 or convolutions and has recently gained popularity for sequential action recognition tasks. Together, these models
 460

⁴⁶⁹ form a diverse evaluation suite for assessing the utility of synthetic IMU data across classical and deep learning action
⁴⁷⁰ recognition classifiers.
⁴⁷¹

⁴⁷² *4.1.6 Configuration.* For the IMU generation stage, the sampling frequency of the IMU data from the physics-based
⁴⁷³ simulation is set to match the frame rate of the source video. Our U-Net model is trained for 100 epochs with a learning
⁴⁷⁴ rate of 1×10^{-3} for both datasets. To evaluate the quality of our synthetic data, we use activity recognition as a
⁴⁷⁵ downstream task to measure the performance gap between models trained on our generated data versus real IMU data.
⁴⁷⁶ All neural classifiers in this downstream setting are trained for 50 epochs with a learning rate of 1×10^{-3} .
⁴⁷⁷

⁴⁷⁹ 4.2 Comparison with the Baseline Method on IMU Generation Quality

⁴⁸¹ Table 2. Comparison between PrimeIMU and a re-implemented baseline IMU generator on the UTD-MHAD and MM-Fit datasets.
⁴⁸² Results are reported per axis for accelerometer (X/Y/Z) and gyroscope (X/Y/Z) signals. IMU generation quality is evaluated using four
⁴⁸³ complementary metrics: mean absolute error (MAE), root mean square error (RMSE), coefficient of determination (R^2), and Pearson
⁴⁸⁴ correlation. For MAE and RMSE, lower values indicate better performance (\downarrow), while for R^2 and Pearson correlation, higher values
⁴⁸⁵ indicate better agreement with real sensor measurements (\uparrow).
⁴⁸⁶

⁴⁸⁷ Dataset	⁴⁸⁸ Method	⁴⁸⁹ Metric	⁴⁹⁰ Accel X	⁴⁹¹ Accel Y	⁴⁹² Accel Z	⁴⁹³ Gyro X	⁴⁹⁴ Gyro Y	⁴⁹⁵ Gyro Z
⁴⁹⁶ UTD-MHAD	⁴⁹⁷ Baseline	MAE \downarrow	0.8395	7.6634	4.6674	1.3719	1.4953	2.1858
		RMSE \downarrow	1.0821	8.2838	5.3449	1.9902	2.2392	2.9975
		$R^2 \uparrow$	-670.6101	-65171.0273	-66641.8406	-1.1429	-1.3515	-2.9680
		Pearson \uparrow	-0.1797	0.1374	-0.1966	0.0143	0.0507	-0.1329
	⁴⁹⁸ PrimeIMU (Ours)	MAE \downarrow	0.0306	0.0176	0.0189	3.0549	3.1581	2.9964
		RMSE \downarrow	0.0399	0.0230	0.0262	4.2513	4.0785	3.8244
		$R^2 \uparrow$	0.8706	0.9785	0.9583	0.9842	0.8805	0.8852
		Pearson \uparrow	0.9886	0.9977	0.9916	0.9972	0.9957	0.9983
⁴⁹⁹ MM-Fit	⁵⁰⁰ Baseline	MAE \downarrow	2.1401	8.0887	6.4013	0.6339	0.7826	1.4799
		RMSE \downarrow	3.1713	9.7756	8.3973	0.9037	1.1181	2.2551
		$R^2 \uparrow$	-1754265773	-1540101166	-7826811739	-3045290257	-17280099211	-41605159758
		Pearson \uparrow	-0.0018	0.0016	-0.0032	-0.005	0.0005	-0.0047
	⁵⁰¹ PrimeIMU (Ours)	MAE \downarrow	0.2808	0.5844	0.2074	0.0691	0.0341	0.0495
		RMSE \downarrow	0.3656	0.6386	0.2764	0.0890	0.0442	0.0655
		$R^2 \uparrow$	-0.1784	-6.4791	0.3422	-0.1988	0.5433	-1.8088
		Pearson \uparrow	0.8963	0.9275	0.9333	0.9455	0.9526	0.9200

⁵⁰³ We compare PrimeIMU with a re-implemented kinematics-only baseline on two multimodal benchmarks, UTD-
⁵⁰⁴ MHAD and MM-Fit, to evaluate IMU generation quality under both in-domain and cross-domain conditions. Results
⁵⁰⁵ are reported per axis for accelerometer (X/Y/Z) and gyroscope (X/Y/Z) signals using four complementary metrics:
⁵⁰⁶ mean absolute error (MAE), root mean square error (RMSE), coefficient of determination (R^2), and Pearson
⁵⁰⁷ correlation (Table 2). Together, these metrics jointly assess amplitude accuracy, variance alignment, and temporal fidelity with
⁵⁰⁸ respect to real sensor measurements.
⁵⁰⁹

⁵¹⁰ Across both datasets, PrimeIMU consistently outperforms the baseline by large margins on all correlation-based
⁵¹¹ metrics, while simultaneously reducing absolute errors. In contrast, the baseline frequently exhibits near-zero or
⁵¹² negative Pearson correlations together with extremely negative R^2 values, indicating a failure to capture meaningful
⁵¹³ temporal structure. PrimeIMU, by comparison, produces signals that closely align with real IMU trajectories over time
⁵¹⁴ and explain a substantial portion of signal variance across nearly all axes.
⁵¹⁵

⁵¹⁶ On UTD-MHAD, PrimeIMU achieves near-ceiling temporal fidelity across all six sensor axes. Pearson correlations
⁵¹⁷ improve from baseline values ranging between -0.20 and 0.14 to above 0.99 for both accelerometer and gyroscope
⁵¹⁸

521 channels. Correspondingly, R^2 increases from extremely negative values (down to -6.6×10^4 on accelerometer axes)
 522 to consistently positive values between 0.87 and 0.98, indicating strong variance alignment with real sensor signals.
 523 Absolute errors are also substantially reduced: accelerometer MAE decreases from 0.84–7.66 to below 0.03, and RMSE
 524 from 1.08–8.28 to below 0.04, representing more than an order-of-magnitude reduction on most axes.
 525

526 Results on MM-Fit further highlight the robustness of PrimeIMU under stronger domain shift. MM-Fit involves higher
 527 variability in exercises, sensor placements, and motion intensities, which more closely reflect real-world deployment
 528 conditions. Under this setting, the baseline effectively loses temporal structure, yielding Pearson correlations near
 529 zero (within ± 0.005) and extremely negative R^2 values (as low as -4.2×10^{11}). In contrast, PrimeIMU maintains strong
 530 temporal alignment across all axes, with Pearson correlations consistently between 0.90 and 0.95. Absolute errors are
 531 also sharply reduced: accelerometer MAE drops from 2.14–8.09 to 0.21–0.58, and RMSE from 3.17–9.78 to 0.28–0.64.
 532

533 For accelerometer signals, these improvements indicate that PrimeIMU recovers both realistic temporal dynamics
 534 and accurate amplitude scaling. Across both datasets, accelerometer MAE and RMSE are reduced by factors ranging
 535 from approximately 10 \times to over 30 \times , while Pearson correlations improve from near-zero or negative values to above
 536 0.90. This suggests that translational motion patterns, which are tightly constrained by underlying kinematics and
 537 physics, are effectively captured by the combined physics-guided initialization and learned refinement stages.
 538

539 For gyroscope signals, PrimeIMU likewise achieves strong temporal fidelity. On UTD-MHAD, Pearson correlations
 540 exceed 0.99 across all gyroscope axes, with R^2 values reaching up to 0.98. While absolute gyroscope errors (MAE/RMSE)
 541 do not always decrease monotonically across all axes, the simultaneous presence of near-ceiling correlation and high
 542 variance explanation indicates that remaining discrepancies primarily reflect scale or bias differences rather than errors
 543 in temporal dynamics. Such behavior is consistent with device-specific angular rate scaling and mounting-dependent
 544 effects, and can typically be addressed through lightweight calibration without altering downstream recognition
 545 performance. On MM-Fit, PrimeIMU dominates the baseline on both error-based and correlation-based metrics, with
 546 gyroscope RMSE reduced by up to an order of magnitude (e.g., Gyro-X 0.90 \rightarrow 0.09) and Pearson correlations improving
 547 from near zero to approximately 0.95.
 548

549 Overall, the quantitative results demonstrate that PrimeIMU substantially outperforms a kinematics-only baseline in
 550 generating realistic and sensor-faithful IMU signals. The most pronounced gains occur in correlation-based measures,
 551 where PrimeIMU consistently recovers correct temporal dynamics and variance structure that the baseline fails to
 552 model. Importantly, these improvements persist across datasets with substantial differences in sensing hardware,
 553 activity distributions, and execution styles, confirming that integrating physics-guided simulation with learned high-
 554 frequency refinement yields inertial signals that are not only numerically accurate, but also dynamically meaningful for
 555 downstream applications.
 556

557 4.3 Classification Performance using Generated IMU Data: Baseline vs. Ours

558 We evaluate whether generated IMU signals can effectively support downstream activity recognition and, critically,
 559 whether they reduce the distribution gap between synthetic and real sensor data. Following prior work, we consider
 560 four training–testing configurations: GT \rightarrow GT (training and testing on real IMU), GEN \rightarrow GEN (training and testing on
 561 generated IMU), GT \rightarrow GEN (training on real IMU and testing on generated IMU), and GEN \rightarrow GT (training on generated
 562 IMU and testing on real IMU). Results on UTD-MHAD and MM-Fit are summarized in Tables 3 and 4 across five
 563 representative classifiers, covering both classical models and deep temporal architectures.
 564

565 Across both datasets and all classifiers, PrimeIMU consistently outperforms the kinematics-only baseline in every
 566 configuration involving synthetic data. This trend holds for Random Forest and SVM as well as for DeepConvLSTM
 567

573 Table 3. Classification results on the UTD-MHAD dataset comparing a re-implemented baseline IMU generator and PrimeIMU across
 574 four training–testing configurations: GT→GT (training and testing on real IMU data), GEN→GEN (training and testing on generated
 575 IMU data), GT→GEN (training on real IMU and testing on generated IMU), and GEN→GT (training on generated IMU and testing on
 576 real IMU). Results are reported for five classifiers using Accuracy (Acc) and macro-F1. Higher values indicate better performance (\uparrow).
 577

578 579 Model	580 Train	581 Test	582 Baseline	583 PrimeIMU (Ours)
			584 Acc \uparrow F1 \uparrow	585 Acc \uparrow F1 \uparrow
586 Random Forest	587 GT	588 GT	589 0.8023 0.7962	590 0.8023 0.7962
	591 GEN	592 GEN	593 0.6734 0.6737	594 0.8166 0.8044
	595 GT	596 GEN	597 0.0201 0.0041	598 0.7880 0.7813
	599 GEN	600 GT	601 0.0516 0.0223	602 0.7937 0.7847
603 SVM	604 GT	605 GT	606 0.7106 0.6908	607 0.7106 0.6908
	608 GEN	609 GEN	610 0.4785 0.4711	611 0.7106 0.6882
	612 GT	613 GEN	614 0.0602 0.0042	615 0.7221 0.7052
	616 GEN	617 GT	618 0.0287 0.0021	619 0.6991 0.6747
620 DeepConvLSTM	621 GT	622 GT	623 0.7564 0.7612	624 0.7564 0.7612
	625 GEN	626 GEN	627 0.4871 0.4793	628 0.7421 0.7303
	629 GT	630 GEN	631 0.0372 0.0057	632 0.7564 0.7537
	633 GEN	634 GT	635 0.0201 0.0124	636 0.7278 0.7119
637 DeepConvLSTM_Attention	638 GT	639 GT	640 0.8940 0.8907	641 0.8940 0.8907
	642 GEN	643 GEN	644 0.7736 0.7744	645 0.9083 0.9033
	646 GT	647 GEN	648 0.0315 0.0024	649 0.8854 0.8750
	650 GEN	651 GT	652 0.0401 0.0168	653 0.8863 0.8813
654 Transformer	655 GT	656 GT	657 0.8625 0.8624	658 0.8625 0.8624
	659 GEN	660 GEN	661 0.7679 0.7717	662 0.8625 0.8563
	663 GT	664 GEN	665 0.0315 0.0023	666 0.8453 0.8396
	667 GEN	668 GT	669 0.0659 0.0236	670 0.8453 0.8375

600
 601
 602 variants and Transformer-based models, indicating that the gains are not tied to a specific classifier architecture
 603 but instead arise from improved signal fidelity at the representation level. In particular, PrimeIMU yields strong
 604 and symmetric performance in both GT→GEN and GEN→GT settings, suggesting a substantial reduction in the
 605 synthetic–real domain gap.
 606

607 On UTD-MHAD, PrimeIMU enables synthetic-only training to closely approach the real-data reference. In the
 608 deployment-relevant GEN→GT setting, the baseline collapses across all models, yielding near-random accuracy and
 609 macro-F1 (e.g., Random Forest 0.0516/0.0223, SVM 0.0287/0.0021, Transformer 0.0659/0.0236). In contrast, PrimeIMU
 610 raises GEN→GT performance to levels comparable to GT→GT. For example, the Transformer achieves 0.8453/0.8375
 611 under GEN→GT compared to 0.8625/0.8624 under GT→GT, while DeepConvLSTM_Attention reaches 0.8863/0.8813
 612 versus 0.8940/0.8907. The remaining gap is consistently small, typically within a few percentage points, indicating that
 613 PrimeIMU preserves most of the discriminative structure required for recognition even when no real IMU data are used
 614 during training.
 615

616 The reverse direction, GT→GEN, provides a complementary assessment of realism. If synthetic signals resemble real
 617 IMU in task-relevant ways, models trained on real data should retain performance when evaluated on synthetic inputs.
 618 This is not the case for the baseline, where GT→GEN accuracy is near zero across all classifiers. PrimeIMU, however,
 619 yields GT→GEN results close to GT→GT, such as 0.8453/0.8396 for the Transformer and 0.7880/0.7813 for Random
 620 Forest. Together with the GEN→GT results, this bidirectional transfer indicates that PrimeIMU substantially reduces
 621

625 Table 4. Classification results on the MM-Fit dataset comparing a re-implemented baseline IMU generator and PrimeIMU across four
 626 training–testing configurations: GT→GT (training and testing on real IMU data), GEN→GEN (training and testing on generated IMU
 627 data), GT→GEN (training on real IMU and testing on generated IMU), and GEN→GT (training on generated IMU and testing on real
 628 IMU). Results are reported for five classifiers using Accuracy (Acc) and macro-F1. Higher values indicate better performance (↑).

Model	Train	Test	Baseline	PrimeIMU (Ours)
			Acc ↑ F1 ↑	Acc ↑ F1 ↑
Random Forest	GT	GT	0.4034 0.3833	0.4034 0.3833
	GEN	GEN	0.2149 0.1613	0.3556 0.3345
	GT	GEN	0.1596 0.1287	0.3499 0.3073
	GEN	GT	0.1664 0.1446	0.3614 0.3342
SVM	GT	GT	0.3572 0.3106	0.3572 0.3106
	GEN	GEN	0.1166 0.0729	0.3512 0.3522
	GT	GEN	0.1088 0.1483	0.3103 0.3013
	GEN	GT	0.1357 0.1138	0.3492 0.3086
DeepConvLSTM	GT	GT	0.3686 0.3488	0.3686 0.3488
	GEN	GEN	0.2461 0.1890	0.3429 0.3063
	GT	GEN	0.1437 0.1140	0.3218 0.2777
	GEN	GT	0.0988 0.0770	0.3322 0.3031
DeepConvLSTM_Attention	GT	GT	0.3541 0.3374	0.3541 0.3374
	GEN	GEN	0.2851 0.2176	0.3446 0.3044
	GT	GEN	0.1587 0.1485	0.2993 0.2667
	GEN	GT	0.1624 0.1267	0.3144 0.2957
Transformer	GT	GT	0.3787 0.3631	0.3787 0.3631
	GEN	GEN	0.2871 0.2106	0.3442 0.2926
	GT	GEN	0.0743 0.0759	0.3086 0.2719
	GEN	GT	0.1376 0.1143	0.3255 0.3030

655 the distribution mismatch between synthetic and real IMU domains in a manner that directly benefits downstream
 656 recognition.

657 Within-domain learnability is reflected by the GEN→GEN setting. PrimeIMU improves GEN→GEN performance over
 658 the baseline across all models, often by large margins (e.g., DeepConvLSTM 0.4871/0.4793 → 0.7421/0.7303, Random
 659 Forest 0.6734/0.6737 → 0.8166/0.8044). Notably, DeepConvLSTM_Attention trained and tested on PrimeIMU signals
 660 slightly exceeds its GT→GT reference. This behavior likely reflects increased temporal regularity in the generated
 661 signals, which can simplify optimization in a fully synthetic setting. Importantly, this effect is confined to within-domain
 662 evaluation and does not suggest that synthetic IMU surpasses real sensor data for real-world deployment.

663 On MM-Fit, absolute accuracies are lower for all methods, consistent with the dataset’s greater variability in exercises,
 664 devices, and sensor placements. This setting more closely reflects practical deployment scenarios, where heterogeneity is
 665 unavoidable. Despite this increased difficulty, PrimeIMU consistently outperforms the baseline across all configurations
 666 and classifiers. In the GEN→GT setting, PrimeIMU improves accuracy by approximately 15–23 points across models
 667 (e.g., Transformer 0.1376 → 0.3255, DeepConvLSTM 0.0988 → 0.3322, SVM 0.1357 → 0.3492), with corresponding
 668 macro-F1 gains of similar magnitude. In the GT→GEN setting, PrimeIMU also yields substantial improvements (e.g.,
 669 Transformer 0.0743/0.0759 → 0.3086/0.2719), indicating that real-trained classifiers remain effective when evaluated
 670 on PrimeIMU-generated signals. GEN→GEN results further confirm that the synthetic IMU distribution produced by
 671 PrimeIMU is internally coherent and learnable, whereas the baseline remains far from competitive.

Overall, these results demonstrate that PrimeIMU-generated IMU signals are not only learnable within the synthetic domain, but also transferable across synthetic and real domains in both directions. The improvements are most pronounced in the deployment-relevant synthetic-only training scenario, where PrimeIMU enables classifiers to generalize effectively to real sensor data without access to real IMU during training. This substantially reduces reliance on large-scale labeled IMU collection and supports the practical use of PrimeIMU for scalable IMU-based activity recognition.

4.4 Data Augmentation

We study whether PrimeIMU can serve as an effective augmentation source when combined with limited real IMU data. Table 5 reports action recognition results on UTD-MHAD when models are trained on real IMU alone (GT) versus a mixture of real and synthetic data (GT+GEN), and evaluated on real IMU.

Table 5. Data augmentation results on the UTD-MHAD dataset. Models are trained either on real IMU data only (GT) or on a combination of real and PrimeIMU-generated synthetic data (GT+GEN), and evaluated on real IMU signals. We report classification accuracy and macro F1 across classical and deep sequential models, highlighting the impact of synthetic IMU augmentation on recognition performance.

Model	Training	Testing	Accuracy ↑	F1 Score (Macro) ↑
RandomForest	GT	GT	0.8023	0.7962
	GT+GEN	GT	0.8381	0.8266
SVM	GT	GT	0.7106	0.6908
	GT+GEN	GT	0.7550	0.7413
DeepConvLSTM	GT	GT	0.7564	0.7612
	GT+GEN	GT	0.8840	0.8743
DeepConvLSTM_Attention	GT	GT	0.8940	0.8907
	GT+GEN	GT	0.9284	0.9247
Transformer	GT	GT	0.8625	0.8624
	GT+GEN	GT	0.8467	0.8483

Across four out of five models, augmenting real training data with PrimeIMU-generated signals consistently improves performance. The gains range from modest improvements for classical classifiers to substantial boosts for deep temporal models, indicating that synthetic IMU signals provide complementary information beyond what is available in the original dataset.

For classical models, augmentation yields stable and meaningful improvements. RandomForest accuracy increases from 0.8023 to 0.8381 (+3.6 points), with macro F1 rising from 0.7962 to 0.8266. SVM shows a similar pattern, improving from 0.7106 to 0.7550 accuracy (+4.4 points) and from 0.6908 to 0.7413 macro F1. These gains suggest that PrimeIMU expands the effective coverage of the feature space, even for shallow models that rely on handcrafted or aggregated temporal features.

The impact is more pronounced for deep recurrent architectures. DeepConvLSTM benefits substantially from augmentation, with accuracy improving from 0.7564 to 0.8840 (+12.8 points) and macro F1 from 0.7612 to 0.8743. DeepConvLSTM_Attention also sees consistent gains, increasing accuracy from 0.8940 to 0.9284 (+3.4 points) and macro F1 from 0.8907 to 0.9247. These results indicate that recurrent temporal models are particularly effective at exploiting the additional fine-grained motion patterns introduced by synthetic IMUs, translating augmented temporal diversity into better generalization on real data.

In contrast, the Transformer shows a small decrease when synthetic data is added ($0.8625 \rightarrow 0.8467$ accuracy). Given its already strong baseline performance, this result suggests that naive augmentation may introduce distributional variability that is not uniformly beneficial for globally attentive architectures. Importantly, this effect is limited to a single model and does not negate the overall trend observed across the remaining classifiers.

Taken together, these findings demonstrate that PrimeIMU is not only suitable for synthetic-only training, but also serves as a practical augmentation mechanism when some real IMU data are available. Augmentation consistently improves or maintains performance for most models, with particularly large gains for recurrent temporal architectures. This suggests that PrimeIMU-generated signals enrich the training distribution in a way that complements real sensor measurements, helping models learn more robust motion representations under limited data regimes.

4.5 Evaluation of Unseen Activity Generation Quality

We further evaluate the generalization ability of PrimeIMU under activity-level distribution shift, focusing on whether realistic IMU signals can be synthesized for activities that are entirely absent during training. This setting reflects a practical deployment scenario, where exhaustive coverage of all activities is infeasible and the model must extrapolate to new motion patterns. To simulate this condition, we progressively remove increasing numbers of activity classes from the training set and evaluate generation quality on the held-out classes. Specifically, we consider four configurations where 1, 3, 7, or 13 activity classes are excluded during training. Performance is assessed using MAE, RMSE, coefficient of determination (R^2), and Pearson correlation across all accelerometer and gyroscope axes. Results are summarized in Table 6.

Generation quality degrades gradually as more activity classes are withheld, reflecting the increasing difficulty of extrapolating to unseen motion patterns. Importantly, this degradation is not catastrophic. Even in the most challenging setting, where 13 out of 27 activities are removed during training, PrimeIMU preserves strong temporal alignment with real sensor signals, as indicated by consistently high Pearson correlations, typically in the range of 0.95–0.98 across both accelerometer and gyroscope channels. High Pearson correlation in this context is particularly important, as it indicates that the temporal evolution and phase structure of motion dynamics are preserved under activity-level distribution shift, even when absolute signal magnitudes vary.

Accelerometer signals exhibit strong robustness to unseen activities. When only a small number of classes are removed (1 or 3), MAE and RMSE remain very low (approximately 0.02–0.04), and Pearson correlations exceed 0.92 across all axes. Even under the most severe removal of 13 activity classes, accelerometer correlations remain high (0.96–0.98), with MAE below 0.05 and largely positive R^2 values. This robustness suggests that translational motion patterns, which are closely tied to underlying kinematic trajectories and gravity constraints, generalize well across activities and are effectively captured by PrimeIMU’s physics-guided simulation and refinement pipeline.

Gyroscope signals show greater sensitivity to unseen activities, particularly as the number of removed classes increases. With only one class excluded, gyroscope channels already achieve strong agreement with real data, with Pearson correlations above 0.92 and positive R^2 across all axes. As more activities are removed, MAE and RMSE increase noticeably, and R^2 becomes negative on several axes (e.g., Gyro-Z under 7- and 13-class removal). These negative R^2 values primarily reflect scale and variance mismatch rather than loss of temporal correspondence, as temporal correlations remain consistently high (0.92–0.98). This behavior is consistent with the fact that angular velocity magnitudes are strongly coupled to activity-specific execution speed, rotation range, and articulation style, and are therefore less directly constrained by kinematic trajectories alone.

781 Table 6. Unseen activity generation performance of PrimeIMU under increasing activity-level distribution shift. We progressively
 782 remove different numbers of activity classes during training (1, 3, 7, and 13 classes) and evaluate IMU generation quality on the held-
 783 out activities. Results are reported per axis for accelerometer and gyroscope signals using MAE, RMSE, coefficient of determination
 784 (R^2), and Pearson correlation, capturing both magnitude accuracy and temporal alignment.

Configuration	Dimension	MAE↓	RMSE ↓	$R^2 \uparrow$	Pearson ↑
w/o 1 class	Accel-X	0.0278	0.0318	-4.5762	0.8223
	Accel-Y	0.0319	0.0347	0.5526	0.9680
	Accel-Z	0.0178	0.0351	-0.9815	0.7919
	Gyro-X	3.1941	4.1490	0.8473	0.9549
	Gyro-Y	2.1869	2.7867	0.4120	0.9200
	Gyro-Z	3.3438	4.5424	0.5921	0.9442
w/o 3 classes	Accel-X	0.0200	0.0249	0.3460	0.9206
	Accel-Y	0.0152	0.0215	0.7880	0.9614
	Accel-Z	0.0206	0.0253	0.3325	0.9298
	Gyro-X	6.3002	6.7170	0.5306	0.9809
	Gyro-Y	6.1725	6.8002	-3.6676	0.8741
	Gyro-Z	3.6351	4.5474	0.2053	0.9547
w/o 7 classes	Accel-X	0.0470	0.0639	-0.6917	0.8887
	Accel-Y	0.0321	0.0457	0.6769	0.9643
	Accel-Z	0.0328	0.0440	0.3705	0.9258
	Gyro-X	5.5519	7.3369	0.4505	0.9772
	Gyro-Y	4.5878	5.9229	-0.7792	0.9254
	Gyro-Z	9.1852	10.4794	-0.7322	0.9777
w/o 13 classes	Accel-X	0.0489	0.0651	0.2695	0.9594
	Accel-Y	0.0501	0.0647	0.5779	0.9779
	Accel-Z	0.0466	0.0592	0.1732	0.9558
	Gyro-X	9.3833	11.8290	-0.8209	0.9755
	Gyro-Y	7.5312	9.5950	-0.6213	0.9545
	Gyro-Z	9.6361	11.5002	-1.4849	0.9751

811
 812 Overall, the unseen-activity evaluation highlights three key findings. First, PrimeIMU maintains strong temporal
 813 fidelity for both accelerometer and gyroscope signals even when synthesizing IMU data for activities that are entirely
 814 unseen during training. Second, accelerometer magnitudes generalize robustly across activities, while gyroscope
 815 channels exhibit larger amplitude deviations under heavy activity removal, suggesting that lightweight calibration may
 816 further improve fidelity in extreme cases. Third, the absence of catastrophic failure under large activity-level distribution
 817 shifts indicates that PrimeIMU learns transferable inertial primitives and motion dynamics rather than memorizing
 818 activity-specific signal templates. This property is essential for scalable IMU synthesis in real-world settings, where
 819 exhaustive activity coverage during training is impractical.
 820
 821
 822

823 4.6 Cross-Dataset Transfer Learning for IMU Generation and Classification

824 A central requirement for practical IMU synthesis is robust cross-dataset generalization: models trained on one dataset
 825 should adapt to new domains with different devices, sampling rates, sensor characteristics, and activity styles using
 826 minimal target-domain data. We evaluate this setting by pretraining PrimeIMU on UTD-MHAD and transferring it
 827 to MM-Fit with 10%, 20%, 30%, and 40% of target-domain data. Signal-level IMU generation quality is reported in
 828 Table 7, while downstream action recognition performance is summarized in Table 8. Together, these experiments
 829
 830
 831
 832 Manuscript submitted to ACM

833 Table 7. Cross-dataset IMU generation quality when transferring from UTD-MHAD to MM-Fit. Models are pretrained on UTD-MHAD
 834 and fine-tuned with different proportions of MM-Fit data (10%, 20%, 30%, 40%). We report signal-level metrics across six sensor
 835 dimensions: mean absolute error (MAE), root mean squared error (RMSE), coefficient of determination (R^2), and Pearson correlation.
 836 These generated signals serve as the basis for the classification experiments in Table 8.

Dimension	Data Proportion	MAE ↓	RMSE ↓	R^2 ↑	Pearson ↑
Accel-X	10%	<u>0.1448</u>	<u>0.1934</u>	0.4315	0.9261
	20%	<u>0.1515</u>	<u>0.1993</u>	0.7276	0.9571
	30%	<u>0.1477</u>	<u>0.1946</u>	0.4309	0.9247
	40%	0.1294	0.1797	<u>0.6439</u>	<u>0.9416</u>
Accel-Y	10%	0.4493	0.4911	-5.5504	0.8799
	20%	<u>0.5704</u>	<u>0.7809</u>	<u>-0.0863</u>	0.6772
	30%	<u>0.2259</u>	<u>0.2699</u>	0.2087	0.9511
	40%	0.1935	0.2626	-0.3392	<u>0.9116</u>
Accel-Z	10%	0.2059	0.2499	-0.0960	0.9444
	20%	0.1146	0.1451	<u>0.5209</u>	0.9743
	30%	0.1904	0.2321	0.3915	0.9528
	40%	<u>0.1191</u>	<u>0.1619</u>	0.7264	<u>0.9533</u>
Gyro-X	10%	<u>0.0500</u>	<u>0.0634</u>	<u>0.2247</u>	<u>0.9760</u>
	20%	0.0342	0.0418	0.7645	0.9827
	30%	0.0558	0.0679	-0.6523	0.9575
	40%	0.0667	0.0783	0.0155	0.9747
Gyro-Y	10%	0.0801	0.0863	-9.4137	0.9598
	20%	0.0266	0.0318	<u>0.2922</u>	<u>0.9738</u>
	30%	<u>0.0442</u>	<u>0.0506</u>	-0.3174	0.9728
	40%	0.0288	0.0352	0.5606	0.9787
Gyro-Z	10%	0.0538	0.0675	-1.5817	0.9440
	20%	0.0446	0.0513	0.6121	0.9746
	30%	<u>0.0531</u>	<u>0.0585</u>	<u>0.1843</u>	<u>0.9622</u>
	40%	0.0591	0.0680	-5.4008	0.9633

865
 866 assess whether PrimeIMU preserves both physically meaningful inertial structure and task-relevant discriminative
 867 information under substantial domain shift.

870 4.6.1 *Signal-Level Cross-Dataset Adaptation.* We first analyze cross-dataset generalization at the signal level using
 871 Table 7, which reports IMU generation quality when PrimeIMU is pretrained on UTD-MHAD and fine-tuned on MM-Fit
 872 with increasing proportions of target-domain data. This analysis isolates whether inertial structure learned from one
 873 dataset can be transferred to another dataset with distinct sensor hardware, sampling frequencies, exercise protocols,
 874 and subject execution styles.

875 With only 10% of MM-Fit data, PrimeIMU already demonstrates strong temporal alignment with real IMU signals.
 876 Pearson correlations exceed 0.92 on most axes, including Accel-X (0.926), Accel-Z (0.944), Gyro-X (0.976), and Gyro-Y
 877 (0.960). This indicates that the temporal evolution and relative phase structure of inertial motion are largely dataset-
 878 invariant and can be transferred from UTD-MHAD to MM-Fit with minimal supervision. At the same time, variance-
 879 based metrics such as R^2 remain low or negative on several axes (e.g., Accel-Y -5.55, Gyro-Y -9.41), reflecting residual
 880 discrepancies in signal magnitude and sensor-specific scaling. The coexistence of high Pearson correlation and low
 881

⁸⁸⁵ R^2 suggests that PrimeIMU captures correct motion dynamics early, while absolute amplitude alignment to the target
⁸⁸⁶ device is not yet fully calibrated.
⁸⁸⁷

⁸⁸⁸ Increasing the fine-tuning ratio to 20% yields clear and consistent improvements across both accelerometer and
⁸⁸⁹ gyroscope channels. For example, Accel-Z MAE decreases from 0.206 to 0.115 and RMSE from 0.250 to 0.145, while
⁸⁹⁰ R^2 becomes positive (0.52) and Pearson increases to 0.974. Similarly, Gyro-X improves to MAE 0.034 with $R^2 = 0.76$
⁸⁹¹ and Pearson 0.983. These results demonstrate that a modest amount of target-domain data is sufficient to substantially
⁸⁹² correct magnitude-related discrepancies on top of already well-aligned temporal structure. The improvements are
⁸⁹³ axis-dependent: accelerometer channels tend to stabilize earlier, whereas gyroscope channels—more sensitive to angular
⁸⁹⁴ velocity scaling, sensor placement, and exercise execution—require additional supervision to achieve reliable variance
⁸⁹⁵ alignment.
⁸⁹⁶

⁸⁹⁷ Beyond 20% fine-tuning, further increases in target-domain data do not lead to consistent gains in signal quality.
⁸⁹⁸ Pearson correlations remain stably high across all axes, typically in the range of 0.95–0.98, indicating that temporal
⁸⁹⁹ motion dynamics are preserved throughout. In contrast, variance-based metrics exhibit non-monotonic behavior,
⁹⁰⁰ particularly for gyroscope channels. For instance, Gyro-Z R^2 decreases from 0.18 at 30% to −5.40 at 40%, despite nearly
⁹⁰¹ unchanged Pearson correlation. We attribute this mild non-monotonicity to the interaction between limited target-
⁹⁰² domain scale diversity and device-specific normalization, rather than to changes in the underlying temporal modeling.
⁹⁰³ Importantly, the stability of Pearson correlation across all fine-tuning ratios confirms that PrimeIMU maintains realistic
⁹⁰⁴ motion dynamics even when exact amplitude alignment varies.
⁹⁰⁵

⁹⁰⁶ Overall, Table 7 reveals a consistent pattern: cross-dataset transfer of temporal inertial structure is highly data-
⁹⁰⁷ efficient, with strong temporal fidelity achieved using as little as 10–20% target-domain data. Magnitude calibration
⁹⁰⁸ improves rapidly at early stages but does not monotonically benefit from additional fine-tuning. This separation between
⁹⁰⁹ temporal fidelity and amplitude alignment is particularly relevant for downstream recognition tasks, which depend
⁹¹⁰ more strongly on relative motion dynamics and inter-axis coordination than on exact signal scaling.
⁹¹¹

⁹¹² **4.6.2 Downstream Classification under Cross-Dataset Transfer.** We next examine whether the signal-level cross-dataset
⁹¹³ generalization observed in Table 7 translates into downstream utility for action recognition. Table 8 reports classification
⁹¹⁴ performance on MM-Fit under four training–testing configurations using PrimeIMU-generated signals at different
⁹¹⁵ fine-tuning ratios.
⁹¹⁶

⁹¹⁷ The most practically relevant setting is GEN→GT, where classifiers are trained entirely on synthetic IMU signals
⁹¹⁸ and evaluated on real MM-Fit data. Across all model families, performance under this setting consistently approaches
⁹¹⁹ the real-data reference (GT→GT), even with limited target-domain fine-tuning. For example, at 20% fine-tuning,
⁹²⁰ the Transformer achieves an accuracy/F1 of 0.331/0.306 under GEN→GT, compared to 0.379/0.351 under GT→GT.
⁹²¹ DeepConvLSTM exhibits a similar trend, reaching 0.334/0.312 versus 0.369/0.349. Classical models follow the same
⁹²² pattern: SVM maintains GEN→GT accuracy around 0.35 across all fine-tuning ratios, only marginally below its GT→GT
⁹²³ performance of 0.357. These results indicate that PrimeIMU preserves the discriminative temporal patterns required for
⁹²⁴ real-world deployment, even when no real IMU data are used during classifier training.
⁹²⁵

⁹²⁶ The GEN→GEN setting provides complementary evidence of internal consistency within the synthetic domain.
⁹²⁷ Models trained and tested entirely on PrimeIMU-generated signals achieve stable and competitive performance across
⁹²⁸ fine-tuning ratios. Notably, SVM reaches 0.353/0.353 at 30% fine-tuning, approaching its GT→GT baseline. Similar
⁹²⁹ trends are observed for Random Forest and deep sequence models, demonstrating that the synthetic IMU distribution is
⁹³⁰ not only realistic relative to real sensors but also coherent and learnable in its own right.
⁹³¹

937 Table 8. Action recognition on MM-Fit using synthetic IMU signals generated in Table 7. We evaluate four train/test configurations: (1)
 938 GT→GT, training and testing on ground-truth IMU; (2) GEN→GEN, training and testing on PrimeIMU-generated IMU; (3) GT→GEN,
 939 trained on real and tested on synthetic; and (4) GEN→GT, trained on synthetic and tested on real. Results are reported under different
 940 fine-tuning proportions of MM-Fit (10%, 20%, 30%, 40%).

Model	Train	Test	10%		20%		30%		40%	
			Acc ↑	F1 ↑						
Random Forest	GT	GT	0.4034	0.3833	0.4034	0.3833	0.4034	0.3833	0.4034	0.3833
	GEN	GEN	0.3490	0.3241	0.3508	0.3278	0.3481	0.3268	0.3495	0.3255
	GT	GEN	0.3582	0.3176	0.3569	0.3182	0.3591	0.3222	0.3675	0.3289
	GEN	GT	0.3663	0.3404	0.3625	0.3382	0.3681	0.3455	0.3711	0.3477
SVM	GT	GT	0.3572	0.3106	0.3572	0.3106	0.3572	0.3106	0.3572	0.3106
	GEN	GEN	0.3508	0.3518	0.3512	0.3531	0.3530	0.3533	0.3516	0.3525
	GT	GEN	0.3112	0.3020	0.3081	0.2989	0.3081	0.2990	0.3055	0.2917
	GEN	GT	0.3505	0.3084	0.3503	0.3086	0.3524	0.3116	0.3524	0.3110
DeepConvLSTM	GT	GT	0.3686	0.3488	0.3686	0.3488	0.3684	0.3423	0.3684	0.3423
	GEN	GEN	0.3556	0.3009	0.3424	0.3134	0.3530	0.3003	0.3235	0.2805
	GT	GEN	0.3174	0.2708	0.3213	0.2792	0.3363	0.2866	0.3275	0.2800
	GEN	GT	0.3427	0.3177	0.3337	0.3123	0.3604	0.3328	0.3461	0.3192
DeepConvLSTM_Attention	GT	GT	0.3541	0.3374	0.3642	0.3440	0.3568	0.3405	0.3568	0.3405
	GEN	GEN	0.3481	0.3103	0.3231	0.2775	0.3429	0.2834	0.3371	0.3023
	GT	GEN	0.3160	0.2724	0.3156	0.2736	0.3182	0.2733	0.3191	0.2777
	GEN	GT	0.3385	0.3151	0.3274	0.3058	0.3335	0.3127	0.3293	0.3052
Transformer	GT	GT	0.3787	0.3631	0.3787	0.3507	0.3740	0.3507	0.3740	0.3507
	GEN	GEN	0.3240	0.2694	0.3490	0.3089	0.3389	0.2849	0.3411	0.2998
	GT	GEN	0.3310	0.2813	0.3279	0.2716	0.3349	0.2847	0.3292	0.2806
	GEN	GT	0.3314	0.3014	0.3312	0.3059	0.3448	0.3168	0.3280	0.3027

964
 965
 966
 967
 968
 969 The GT→GEN configuration evaluates whether synthetic signals are perceived as realistic by models trained on
 970 real IMU data. Performance remains stable across fine-tuning ratios, with moderate and controlled gaps relative to
 971 GT→GT. For instance, the Transformer achieves 0.335/0.285 at 30% fine-tuning, compared to 0.379/0.363 under GT→GT.
 972 Unlike prior kinematics-only synthesis approaches, where GT-trained models often degrade sharply on synthetic inputs,
 973 PrimeIMU-generated signals remain compatible with real-trained classifiers. This behavior aligns with the signal-level
 974 observation that temporal motion dynamics are faithfully preserved across domains.
 975

976 Across all classifiers, classification performance improves rapidly between 10% and 20% fine-tuning and then stabilizes,
 977 mirroring the saturation behavior observed in signal-level Pearson correlation and temporal metrics. This consistency
 978 suggests that downstream recognition models primarily rely on relative temporal structure and inter-axis coordination,
 979 which are already well-aligned at low fine-tuning ratios. Consequently, exact amplitude calibration, while beneficial for
 980 signal fidelity, plays a secondary role in enabling cross-dataset transfer for IMU-based activity recognition.
 981

982 Taken together, the results demonstrate that PrimeIMU enables effective cross-dataset transfer at both the signal and
 983 task levels. With limited target-domain data, the framework captures dataset-invariant inertial dynamics that support
 984 robust action recognition across heterogeneous sensor configurations, substantially reducing the need for large-scale
 985 recollection of labeled IMU data when deploying recognition systems in new environments.
 986

989 Table 9. Ablation study of PrimeIMU on the UTD-MHAD dataset. We compare the full model with variants where key components
 990 are removed: pose sequence input (w/o pose), simulated IMU initialization from the physics stage (w/o phy imu), and individual loss
 991 terms for kinematic consistency ($\mathcal{L}_{\text{kine}}$), spectral alignment ($\mathcal{L}_{\text{stft}}$), and time-domain reconstruction ($\mathcal{L}_{\text{recon}}$). Results are reported
 992 across six sensor dimensions (Accel-X/Y/Z, Gyro-X/Y/Z) using MAE, RMSE, R^2 , and Pearson correlation. Bold denotes the best result
 993 within each sensor dimension, and underlines indicate the second-best.

Dimension	Model	MAE ↓	RMSE ↓	$R^2 \uparrow$	Pearson ↑
Accel-X	Full	0.0306	0.0399	0.8706	0.9886
	w/o pose	0.0235	0.0319	0.9323	<u>0.9811</u>
	w/o phy imu	0.0722	0.1118	0.3557	0.8347
	w/o $\mathcal{L}_{\text{kine}}$	<u>0.0295</u>	<u>0.0396</u>	<u>0.8902</u>	0.9762
	w/o $\mathcal{L}_{\text{stft}}$	0.0745	0.1117	0.3190	0.9494
	w/o $\mathcal{L}_{\text{recon}}$	0.0401	0.0520	0.1943	0.9708
Accel-Y	Full	0.0176	0.0230	0.9785	0.9977
	w/o pose	<u>0.0234</u>	<u>0.0326</u>	0.9585	0.9900
	w/o phy imu	0.0285	0.0384	0.9679	0.9933
	w/o $\mathcal{L}_{\text{kine}}$	0.0275	0.0373	0.9721	<u>0.9918</u>
	w/o $\mathcal{L}_{\text{stft}}$	0.0284	0.0366	<u>0.9725</u>	0.9922
	w/o $\mathcal{L}_{\text{recon}}$	0.0374	0.0475	0.9349	0.9912
Accel-Z	Full	0.0189	0.0262	0.9583	0.9916
	w/o pose	0.0276	0.0349	0.9343	<u>0.9911</u>
	w/o phy imu	0.0349	0.0459	0.8822	0.9822
	w/o $\mathcal{L}_{\text{kine}}$	0.0396	0.0496	0.8891	0.9763
	w/o $\mathcal{L}_{\text{stft}}$	<u>0.0250</u>	<u>0.0323</u>	0.9205	0.9884
	w/o $\mathcal{L}_{\text{recon}}$	0.0440	0.0559	0.7398	0.9741
Gyro-X	Full	3.0549	4.2513	0.9842	0.9972
	w/o pose	4.5961	5.7856	0.9462	0.9944
	w/o phy imu	4.8794	6.4089	0.9337	0.9910
	w/o $\mathcal{L}_{\text{kine}}$	<u>4.0421</u>	5.6858	0.9814	0.9926
	w/o $\mathcal{L}_{\text{stft}}$	4.1948	<u>5.5147</u>	<u>0.9750</u>	0.9945
	w/o $\mathcal{L}_{\text{recon}}$	9.4388	11.5457	0.8519	0.9862
Gyro-Y	Full	3.1581	4.0785	0.8006	0.9957
	w/o pose	<u>3.5308</u>	<u>4.2360</u>	0.8109	<u>0.9938</u>
	w/o phy imu	4.9996	6.6176	0.8989	0.9736
	w/o $\mathcal{L}_{\text{kine}}$	4.3202	5.9407	0.6626	0.9528
	w/o $\mathcal{L}_{\text{stft}}$	4.6992	5.9767	0.7854	0.9808
	w/o $\mathcal{L}_{\text{recon}}$	5.5411	7.0502	0.5790	0.9813
Gyro-Z	Full	2.9964	3.8424	0.8852	0.9983
	w/o pose	<u>4.3181</u>	<u>5.1476</u>	0.8355	<u>0.9965</u>
	w/o phy imu	5.2329	6.6543	0.8841	0.9926
	w/o $\mathcal{L}_{\text{kine}}$	4.8615	5.9656	0.8339	0.9935
	w/o $\mathcal{L}_{\text{stft}}$	5.1138	6.4306	0.9557	0.9906
	w/o $\mathcal{L}_{\text{recon}}$	6.9763	8.8849	0.8483	0.9863

4.7 Ablation Analysis

We conduct an ablation study on UTD-MHAD to quantify the contribution of PrimeIMU’s core design components. Specifically, we ablate (i) input sources, including low-frequency pose guidance and physics-inspired simulated IMU initialization, and (ii) objective terms, including time-domain reconstruction loss $\mathcal{L}_{\text{recon}}$, multi-resolution spectral alignment loss $\mathcal{L}_{\text{stft}}$, and kinematic consistency loss $\mathcal{L}_{\text{kine}}$. Performance is evaluated across six sensor dimensions using

1041 MAE and RMSE to assess absolute amplitude fidelity, together with R^2 and Pearson correlation to capture variance
 1042 alignment and temporal agreement. This combination allows us to distinguish errors in signal scale from mismatches
 1043 in motion dynamics.
 1044

1045 *4.7.1 Effect of Input Sources: Pose Guidance and Physics-Based Initialization.* We first examine the role of PrimeIMU's
 1046 two complementary input sources. Removing the physics-based simulated IMU initialization (w/o phy imu) leads
 1047 to a consistent and substantial degradation across both accelerometer and gyroscope channels. For accelerometers,
 1048 errors increase sharply (e.g., Accel-X MAE increases from 0.0306 to 0.0722 and Pearson drops from 0.9886 to 0.8347),
 1049 while gyroscopes exhibit pronounced scale mismatches (e.g., Gyro-X MAE increases from 3.05 to 4.88). These results
 1050 indicate that the physics stage injects essential high-frequency and device-like inertial structure that cannot be reliably
 1051 recovered from pose alone, particularly for angular velocity signals.
 1052

1053 *4.7.2 Effect of Loss Functions: Reconstruction, Spectral Alignment, and Kinematic Consistency.* We next analyze the
 1054 contribution of individual loss terms, each of which targets a distinct failure mode in IMU synthesis. The reconstruction
 1055 loss $\mathcal{L}_{\text{recon}}$ primarily controls amplitude and bias alignment in the time domain. Removing it causes the most severe
 1056 degradation on gyroscopes, with MAE and RMSE increasing dramatically (e.g., Gyro-X MAE 9.44 vs. 3.05; Gyro-Z MAE
 1057 6.98 vs. 3.00) and R^2 dropping substantially, even though Pearson correlation remains relatively high. This reflects
 1058 a characteristic failure mode in which temporal trends are preserved but signal magnitudes drift, underscoring the
 1059 necessity of explicit time-domain supervision for sensor-faithful synthesis.
 1060

1061 Overall, the two inputs serve distinct and complementary roles. Physics-based initialization provides realistic device-
 1062 level dynamics and high-frequency structure, while pose guidance stabilizes global motion trends and gravity orientation.
 1063 Using both jointly yields the strongest and most consistent performance across sensors and metrics.
 1064

1065 *4.7.3 Axis-Wise Behavior and Sensor-Specific Sensitivity.* Examining axis-wise behavior reveals patterns consistent with
 1066 inertial sensor physics. Accelerometer channels benefit most from the full composite objective, achieving near-ceiling
 1067

1068 The spectral alignment loss $\mathcal{L}_{\text{stft}}$ shapes frequency content and suppresses non-physical oscillations introduced by
 1069 differentiation. Without it, accelerometer errors increase markedly (e.g., Accel-X MAE 0.0745 vs. 0.0306; Accel-Z RMSE
 1070 0.0323 vs. 0.0262), accompanied by a clear drop in R^2 . Gyroscope channels also suffer increased scale error and reduced
 1071 correlation (e.g., Gyro-Z MAE 5.11 vs. 3.00). These results confirm that spectral supervision is critical for restoring
 1072 IMU-like frequency characteristics that are not sufficiently constrained by time-domain losses alone.
 1073

1074 The kinematic consistency loss $\mathcal{L}_{\text{kine}}$ enforces agreement between generated signals and pose-derived motion
 1075 after integration. Disabling it disproportionately affects axes sensitive to accumulated drift and gravity alignment. For
 1076 example, Accel-Z RMSE increases from 0.0262 to 0.0496 and Gyro-Z MAE rises from 3.00 to 4.86, with corresponding
 1077 reductions in correlation. This demonstrates that $\mathcal{L}_{\text{kine}}$ acts as an effective physical regularizer, preventing non-physical
 1078 accelerations and long-term drift while allowing the model to learn fine-grained inertial detail.
 1079

1080 *4.7.4 Summary.* We have presented PrimeIMU, a novel framework for high-frequency video-to-IMU synthesis. Our
 1081 contributions include a physics-guided simulation module that generates initial IMU signals, a hybrid U-Net refinement
 1082 module that performs multi-scale feature extraction and synthesis, and a composite loss function that integrates multiple
 1083 metrics to guide the learning process. We demonstrate that PrimeIMU can synthesize high-quality IMU signals from
 1084 low-quality video sequences, achieving performance comparable to state-of-the-art methods. Our analysis shows that
 1085 the physics-based initialization is crucial for maintaining device-like dynamics and high-frequency structure, while the
 1086 spectral supervision is key for restoring IMU-like frequency characteristics. The kinematic consistency loss is effective
 1087 for preventing long-term drift and learning fine-grained inertial detail. Finally, our ablation studies highlight the
 1088 importance of each component in the composite objective. Future work includes extending PrimeIMU to handle
 1089 multi-camera video sequences and exploring its potential applications in robotics and virtual reality.

1093 correlations (e.g., Accel-Y Pearson 0.9977) and high variance explanation. They are particularly sensitive to the removal
1094 of spectral supervision and physics priors, both of which directly influence frequency content and impact responses.
1095

1096 Gyroscope channels, in contrast, maintain very high Pearson correlations across most ablations, indicating robust
1097 temporal alignment. However, they are significantly more sensitive to scale and bias errors when $\mathcal{L}_{\text{recon}}$ or $\mathcal{L}_{\text{stft}}$ is
1098 removed, as reflected by large increases in MAE and RMSE despite stable correlations. This recurring “high-correlation
1099 but wrong-scale” behavior motivates the inclusion of explicit amplitude- and spectrum-alignment terms and highlights
1100 the importance of reporting both correlation-based and error-based metrics.
1101

1102 *4.7.4 Summary and Implications.* While isolated ablations may yield marginal improvements on individual axes or
1103 metrics, the full PrimeIMU configuration consistently provides the strongest joint trade-off across MAE, RMSE, R^2 , and
1104 Pearson over all six sensor channels. The ablation results confirm that each component addresses a complementary
1105 aspect of IMU synthesis: physics priors inject device realism, pose guidance stabilizes global motion, reconstruction
1106 enforces correct scale, spectral alignment restores frequency structure, and kinematic consistency prevents drift.
1107 Together, these components produce sensor-faithful inertial signals that directly support the cross-dataset transfer and
1108 downstream action recognition gains demonstrated in Section 4.
1109

1110 5 Conclusion

1111 We introduce PrimeIMU, a framework for synthesizing high-frequency, high-fidelity IMU signals directly from video.
1112 Our study demonstrates that PrimeIMU addresses three long-standing challenges in video-to-IMU generation: (i)
1113 error propagation from imperfect pose estimation, (ii) the loss of fine-grained motion dynamics caused by low video
1114 frame rates, and (iii) the mismatch between kinematic estimates and the measurements by physical IMU sensors. By
1115 combining physics-guided IMU simulation with a hybrid U-Net refinement module, PrimeIMU first converts video-
1116 derived kinematic trajectories into approximate inertial measurements and then learns to refine these simulated signals
1117 to more faithfully capture the dynamics and sensor characteristics of real IMU devices. Extensive experiments show
1118 that PrimeIMU (1) improves the fidelity of inertial signal synthesis, (2) generalizes to unseen motion patterns while
1119 preserving realistic signal characteristics, (3) enables synthetic-only training of downstream models with performance
1120 approaching that of real-sensor training and (4) provides additional gains when used to augment real data, and (5)
1121 adapts across datasets and sensor configurations with minimal fine-tuning, supporting robust downstream performance
1122 under domain shift.
1123

1145 **References**

- 1146 [1] Ghaith Al-refai, Dina Karasneh, Hisham Elmoaqet, Mutaz Ryalat, and Natheer Almtireen. 2025. Surface Classification from Robot Internal
1147 Measurement Unit Time-Series Data Using Cascaded and Parallel Deep Learning Fusion Models. *Machines* 13, 3 (2025), 251.
- 1148 [2] Lei Bai, Lina Yao, Xianzhi Wang, Salil S Kanhere, Bin Guo, and Zhiwen Yu. 2020. Adversarial multi-view networks for activity recognition.
1149 *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2 (2020), 1–22.
- 1150 [3] Dmitrije Balabka. 2019. Semi-supervised learning for human activity recognition using adversarial autoencoders. In *Adjunct proceedings of the 2019
1151 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2019 ACM international symposium on wearable
1152 computers*. 685–688.
- 1153 [4] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. 2015. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth
1154 camera and a wearable inertial sensor. In *International Conference on Image Processing*.
- 1155 [5] Wenqiang Chen, Shupei Lin, Elizabeth Thompson, and John Stankovic. 2021. Sensecollect: We need efficient ways to collect on-body sensor-based
1156 human activity data! *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–27.
- 1157 [6] Dongzhou Cheng, Lei Zhang, Can Bu, Xing Wang, Hao Wu, and Aiguo Song. 2023. ProtoHAR: Prototype guided personalized federated learning for
1158 human activity recognition. *IEEE Journal of Biomedical and Health Informatics* 27, 8 (2023), 3900–3911.
- 1159 [7] Giovanni Cioffi, Leonard Bauersfeld, Elia Kaufmann, and Davide Scaramuzza. 2023. Learned inertial odometry for autonomous drone racing. *IEEE
1160 Robotics and Automation Letters* 8, 5 (2023), 2684–2691.
- 1161 [8] Asiya Demirtaş, Gökhan Erdemir, and Haluk Bayram. 2024. Indoor surface classification for mobile robots. *PeerJ Computer Science* 10 (2024), e1730.
- 1162 [9] Siwei Feng and Marco F Duarte. 2019. Few-shot learning-based human activity recognition. *Expert Systems with Applications* 138 (2019), 112782.
- 1163 [10] Iuri Frosio, Federico Pedersini, and N Alberto Borghese. 2008. Autocalibration of MEMS accelerometers. *IEEE Transactions on instrumentation and
1164 measurement* 58, 6 (2008), 2034–2041.
- 1165 [11] Walid Gomaa and Mohamed A Khamis. 2023. A perspective on human activity recognition from inertial motion data. *Neural Computing and
1166 Applications* (2023).
- 1167 [12] Harish Haresamudram, Apoorva Beedu, Varun Agrawal, Patrick L Grady, Irfan Essa, Judy Hoffman, and Thomas Plötz. 2020. Masked reconstruction
1168 based self-supervision for human activity recognition. In *Proceedings of the 2020 ACM International Symposium on Wearable Computers*. 45–49.
- 1169 [13] Hyekhyun Kwon, Catherine Tong, Harish Haresamudram, Yan Gao, Gregory D Abowd, Nicholas D Lane, and Thomas Ploetz. 2020. Imutube:
1170 Automatic extraction of virtual on-body accelerometry from video for human activity recognition. *Proceedings of the ACM on Interactive, Mobile,
1171 Wearable and Ubiquitous Technologies* (2020).
- 1172 [14] Artuu Lämsä, Jaakko Tervonen, Jussi Liikka, Constantino Álvarez Casado, and Miguel Bordallo López. 2022. Video2IMU: Realistic IMU features and
1173 signals from videos. In *International Conference on Wearable and Implantable Body Sensor Networks*.
- 1174 [15] Zikang Leng, Amitrajit Bhattacharjee, Hrudhai Rajasekhar, Lizhe Zhang, Elizabeth Bruda, Hyekhyun Kwon, and Thomas Plötz. 2024. Imugpt 2.0:
1175 Language-based cross modality transfer for sensor-based human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and
1176 Ubiquitous Technologies* (2024).
- 1177 [16] Zikang Leng, Hyekhyun Kwon, and Thomas Ploetz. 2023. Generating Virtual On-Body Accelerometer Data from Virtual Textual Descriptions for
1178 Human Activity Recognition. In *International Symposium on Wearable Computers*.
- 1179 [17] Faisal Mohd-Yasin, Can E Korman, and David J Nagel. 2003. Measurement of noise characteristics of MEMS accelerometers. *Solid-State Electronics*
1180 47, 2 (2003), 357–360.
- 1181 [18] Andreas Mueller. 2019. Modern robotics: Mechanics, planning, and control [bookshelf]. *IEEE Control Systems Magazine* 39, 6 (2019), 100–102.
- 1182 [19] Bartłomiej Nalepa, Magdalena Pawlyta, Mateusz Janiak, Agnieszka Szczęsna, Aleksander Gwiazda, and Konrad Wojciechowski. 2022. Recreating
1183 the Motion Trajectory of a System of Articulated Rigid Bodies on the Basis of Incomplete Measurement Information and Unsupervised Learning.
1184 *Sensors* 22, 6 (2022), 2198.
- 1185 [20] Francisco Javier Ordóñez and Daniel Roggen. 2016. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity
1186 recognition. *Sensors* (2016).
- 1187 [21] Aakash Parmar, Rakesh Katariya, and Vatsal Patel. 2018. A review on random forest: An ensemble classifier. In *International Conference on Intelligent
1188 Data Communication Technologies and Internet of Things*.
- 1189 [22] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 2019. 3d human pose estimation in video with temporal convolutions and
1190 semi-supervised training. In *Conference on Computer Vision and Pattern Recognition*.
- 1191 [23] Thomas Plötz. 2023. If only we had more data!: Sensor-based human activity recognition in challenging scenarios. In *2023 IEEE International
1192 Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. IEEE, 565–570.
- 1193 [24] Sathian Pookkuttath, Povendhan Arthanaripalayam Palanisamy, and Mohan Rajesh Elara. 2023. AI-Enabled Condition Monitoring Framework for
1194 Outdoor Mobile Robots Using 3D LiDAR Sensor. *Mathematics* 11, 16 (2023), 3594.
- 1195 [25] Yuantian Qin, Zhehang Yin, Quanou Yang, and Kai Zhang. 2024. Dynamics Parameter Identification of Articulated Robot. *Machines* 12, 9 (2024),
1196 595.
- 1197 [26] Yuheng Qiu, Can Xu, Yutian Chen, Shibo Zhao, Junyi Geng, and Sebastian Scherer. 2025. AirIO: Learning Inertial Odometry with Enhanced IMU
1198 Feature Observability. *arXiv preprint arXiv:2501.15659* (2025).

- 1197 [27] Daniele Ravi, Clarence Wong, Benny Lo, and Guang-Zhong Yang. 2016. Deep learning for human activity recognition: A resource efficient
 1198 implementation on low-power devices. In *2016 IEEE 13th international conference on wearable and implantable body sensor networks (BSN)*. IEEE,
 1199 71–76.
- 1200 [28] Aaqib Saeed, Tanir Ozcelebi, and Johan Lukkien. 2019. Multi-task self-supervised learning for human activity detection. *Proceedings of the ACM on
 1201 Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (2019), 1–30.
- 1202 [29] Panneer Selvam Santhalingam, Parth Pathak, Huzefa Rangwala, and Jana Kosecka. 2023. Synthetic smartwatch imu data generation from in-the-wild
 1203 asl videos. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* (2023).
- 1204 [30] Allah Bux Sargano, Xiaofeng Wang, Plamen Angelov, and Zulfiqar Habib. 2017. Human action recognition using transfer learning with deep
 1205 representations. In *2017 International joint conference on neural networks (IJCNN)*. IEEE, 463–469.
- 1206 [31] Ronald W Schafer. 2011. What is a savitzky-golay filter?[lecture notes]. *IEEE Signal Processing Magazine* (2011).
- 1207 [32] Satya P Singh, Madan Kumar Sharma, Aimé Lay-Ekuakille, Deepak Gangwar, and Sukrit Gupta. 2020. Deep ConvLSTM with self-attention for
 1208 human activity decoding using wearable sensors. *IEEE Sensors Journal* (2020).
- 1209 [33] David Strömbäck, Sangxia Huang, and Valentin Radu. 2020. Mm-fit: Multimodal deep learning for automatic exercise logging across sensing devices.
 1210 *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* (2020).
- 1211 [34] Ruslan Sultan and Steffen Greiser. 2025. Time and Frequency Domain Analysis of IMU-based Orientation Estimation Algorithms with Comparison
 1212 to Robotic Arm Orientation as Reference. *Sensors* 25, 16 (2025), 5161.
- 1213 [35] Shan Suthaharan. 2016. Support vector machine. In *Machine learning models and algorithms for big data classification: thinking with examples for
 1214 effective learning*. Springer.
- 1215 [36] Nilay Tufek, Murat Yalcin, Mucahit Altintas, Fatma Kalaoglu, Yi Li, and Senem Kursun Bahadir. 2019. Human action recognition using deep learning
 1216 methods on limited sensory data. *IEEE Sensors Journal* (2019).
- 1217 [37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is
 1218 all you need. *Advances in Neural Information Processing Systems* (2017).
- 1219 [38] Gregory W Vogl, M Alkan Donmez, Andreas Archenti, and Brian A Weiss. 2016. Inertial measurement unit for on-Machine diagnostics of machine
 1220 tool linear axes. In *Annual Conference of the PHM Society*, Vol. 8.
- 1221 [39] Cong Xu, Yuhang Li, Dae Lee, Dae Hoon Park, Hongda Mao, Huyen Do, Jonathan Chung, and Dinesh Nair. 2023. Augmentation robust self-supervised
 1222 learning for human activity recognition. In *International Conference on Acoustics, Speech and Signal Processing*.
- 1223 [40] Yafei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. 2023. Vitpose++: Vision transformer for generic body pose estimation. *Transactions on
 1224 Pattern Analysis and Machine Intelligence* (2023).
- 1225 [41] Xiyuan Zhang, Diyan Teng, Ranak Roy Chowdhury, Shuheng Li, Dezhi Hong, Rajesh Gupta, and Jingbo Shang. 2024. UniMTS: Unified Pre-training
 1226 for Motion Time Series. In *Advances in Neural Information Processing Systems*.

1227 Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

1228

1229

1230

1231

1232

1233

1234

1235

1236

1237

1238

1239

1240

1241

1242

1243

1244

1245

1246

1247

1248