A Large-Scale Empirical Study of Deep Learning for MEMS Gyroscope Bias Estimation

Chia-Tung Chung[a], Hsiu-Chi Tsai[a,*]

[a]*National Yang Ming Chiao Tung University, Hsinchu, Taiwan*

Abstract

MEMS gyroscope bias drift is a dominant error source in sensor calibration and inertial navigation. No prior work directly compares deep learning architectures for gyroscope bias estimation on common benchmarks. We report 2,811 experiments comparing MLP, LSTM, TCN, and Transformer models on two public visual-inertial datasets (EuRoC, TUM-VI) with leave-one-sequence-out cross-validation and 17 random seeds. Variance decomposition reveals that dataset choice explains ≈97% of performance variation; model architecture accounts for less than 1%. On clean-label EuRoC, TCN achieves the highest attitude-drift improvement (96.19%, Friedman $p<0.001$); on noisy-label TUM-VI, LSTM reaches 21.57%, but only one of six pairwise comparisons is significant. Physics-informed SO(3) regularization fails to improve performance across 390 confirmatory experiments ($N=17$). Removing a common label-denoising step changes EuRoC results by +1.2–3.5 pp and reverses the TUM-VI ranking. A within-dataset label corruption experiment confirms that label quality drives performance far more than architecture selection (≈40:1). Label quality and evaluation protocol matter more than architecture choice here.

Keywords: MEMS gyroscope, sensor calibration, measurement data processing, inertial measurement unit, deep learning, bias estimation

*Corresponding author

Email addresses: jun.514114.ee10@nycu.edu.tw (Chia-Tung Chung), hctsai1006@cs.nctu.edu.tw (Hsiu-Chi Tsai)

URL: https://www.cs.nycu.edu.tw/ (Hsiu-Chi Tsai)

1. Introduction

Microelectromechanical system (MEMS) gyroscopes are the dominant angular-rate sensors in consumer, automotive, and robotics applications, yet their bias drift (which varies with temperature, mechanical stress, and aging) remains a primary error source in inertial navigation and sensor calibration [1, 2]. Traditional calibration methods (multi-position turntable procedures, polynomial temperature models) require controlled laboratory conditions and periodic recalibration, limiting their applicability to field-deployed systems.

Several groups have turned to neural networks that learn to predict gyroscope bias directly from raw IMU signals, with reported accuracies as high as R^2=0.9998 for an MLP [3], compact ∼222-parameter models running on microcontrollers [4], and pre-trained encoders that transfer across sensor platforms [5]. From a sensor-data-processing perspective, these studies are promising, but each evaluates one architecture on one dataset. As a result, it remains unclear whether reported gains stem from the model, the data, or uncontrolled experimental factors.

A recent survey of deep learning for inertial navigation [6] notes the lack of common benchmarks across platforms; we are not aware of any study that compares multiple architectures for IMU bias estimation under controlled conditions. The closest prior work [7] evaluates a single learned-bias model within a factor-graph framework. More broadly, the IMU deep learning literature suffers from three methodological shortcomings: (i) most studies report 1–3 training runs without statistical testing; (ii) random train/validation splits leak temporal structure in time-series data [8, 9]; and (iii) hardware and hyperparameter confounds (GPU type, batch size, learning rate) are rarely controlled.

The absence of controlled multi-architecture comparisons has practical consequences. Practitioners must choose among MLP, LSTM, TCN, and Transformer variants for their specific IMU and deployment constraints, yet available evidence comes from single-architecture papers that differ in datasets, seeds, preprocessing, and GPU hardware. Uncontrolled confounds can shift performance by tens of percentage points (Section 5.5), making cross-paper comparisons unreliable.

We make five contributions:

1. A **unified comparison** of four architectures (MLP, LSTM, TCN, Transformer) on two public datasets (EuRoC, TUM-VI) under identical

training protocols, totaling 2,811 experiments.

2. A **variance decomposition** showing that dataset choice explains ≈97% of performance variance, while model architecture accounts for less than 1%. A single preprocessing decision (label denoising) shifts EuRoC by +1.2–3.5 pp and reverses TUM-VI rankings.
3. A **null result on physics-informed regularization**: SO(3) rotation constraints fail to improve estimation quality in 390 cross-GPU experiments.
4. A **failure-mode analysis** identifying Room1 pathology, temporal-split leakage, and cross-GPU noise amplification as reproducibility hazards.
5. **Practical recommendations** for training protocol, OOD gating, and statistical reporting in IMU bias-estimation studies.

Sections 2–3 cover related work and experimental setup; Sections 4–5 present results and analysis; Section 6 discusses implications.

2. Related Work

2.1. ML for IMU Bias Estimation

Deep learning for inertial sensing spans bias calibration, denoising, and odometry, using LSTMs [10], TCNs [11], and Transformers [12] with residual connections [13] and layer normalization [14]. Brossard et al. [15] denoise gyroscopes via dilated convolutions on EuRoC and TUM-VI; TLIO [16] learns IMU-only odometry; AirIMU [17] corrects inertial measurements for downstream integration; TartanIMU [5] pre-trains on 100+ hours across 10 robotic platforms; TinyGC-Net [4] fits ∼222 parameters on a Cortex-M4; Huang et al. [18] and Calib-Net [19] learn multi-axis calibration; and Long et al. [3] report R^2=0.9998 with an MLP. Newer directions include diffusion-based bias estimation [20], Lie-group dynamics [21], and Transformer-based denoising [22].

LGC-Net [23] also evaluates on EuRoC and TUM-VI but uses a single architecture. We are not aware of prior work that compares MLP, LSTM, TCN, and Transformer on the same bias-estimation task under a unified protocol with repeated seeds and statistical testing. A recent survey [6] corroborates this gap.

2.2. Evaluation Methodology

Time-series evaluation requires temporal splitting to prevent information leakage [8, 9, 24]; we ensure chronological train/validation splits within each LOSO fold. Statistical model comparison [25] is standard in ML but rare in the IMU literature, where most studies report 1–3 runs without confidence intervals [26]. New benchmarks like IF-D [27] are beginning to address this. We use bootstrap BCa intervals [28] and paired permutation tests [29].

2.3. Physics-Informed Regularization

PINNs embed domain constraints as soft penalties but are susceptible to loss-function conflicts [30], spectral bias [31], and precision issues [32]. Conflict-free gradient methods [33] mitigate multi-objective imbalance but have not been applied to IMU bias estimation. Our 390-run PINN study is, to the best of our knowledge, the first to evaluate SO(3) regularization for this task.

3. Experimental Setup

We compare four neural network architectures for MEMS gyroscope bias estimation on two public visual-inertial datasets under leave-one-sequence-out (LOSO) cross-validation.

All experiments share one evaluation protocol, training pipeline, and statistical testing framework.

3.1. Datasets

3.1.1. EuRoC MAV

The EuRoC dataset [34] provides visual-inertial recordings from an AscTec Firefly hexacopter with an ADIS16448 IMU at 200 Hz. Ground-truth bias labels come from batch optimization [35] fusing motion-capture poses with IMU preintegration. We use the five Vicon Room sequences (106,095 samples, $\approx$8.8 min).

3.1.2. TUM-VI

The TUM-VI dataset [36] consists of handheld recordings with a BMI160 IMU at 200 Hz and OptiTrack ground truth at 120 Hz. We select six indoor rooms (161,617 samples, $\approx$13.5 min). Reference angular velocities are derived by differentiating the interpolated quaternion trajectory: $\omega_{\text{ref},k} = 2\, q_k^{-1} \otimes \dot{q}_k$, where $\dot{q}$ is estimated via Savitzky–Golay smoothing (window = 31, polynomial

Table 1: Dataset overview.

Dataset	Platform	Seq.	Rate	Samples
EuRoC	Hexacopter (ADIS16448)	5	200 Hz	106,095
TUM-VI	Handheld (BMI160)	6	200 Hz	161,617

Table 2: Model architectures (18-dim input, 3-axis output).

Model	Params	L	Configuration
MLP	5,571	—	[64, 64], ReLU, dropout 0.2
LSTM	60,548	50	2-layer, h=64, LayerNorm, attn. pool
TCN	241,252	50	5 blocks [32−128], k=3
Transformer	530,947	200	d=128, H=4, 2 layers, [CLS]

order = 3) of the SLERP-interpolated OptiTrack quaternions. Unlike EuRoC, whose labels come from batch optimization that explicitly estimates gyroscope bias, TUM-VI requires *pseudo-bias* labels: $b_k = \omega_{\text{meas},k} - \omega_{\text{ref},k}$. These labels inherit noise from both gyroscope angle random walk and quaternion-differentiation uncertainty. Empirically, the label standard deviation ($\sigma \approx 1.4\,\text{rad}\,\text{s}^{-1}$, estimated as the RMS of b_k after subtracting a 10-second moving average) exceeds the true bias magnitude ($\sim 0.01\,\text{rad}\,\text{s}^{-1}$) by two orders of magnitude, explaining the EuRoC–TUM-VI performance gap (Section 4).

Room1 has only $\approx 4^\circ$ baseline drift, making percentage metrics unstable (Section 4.2).

3.2. Features and Models

For each timestep, we extract an 18-dimensional feature vector from a sliding window of W=50 samples (0.25 s): instantaneous gyroscope (3 D), gyroscope window mean and standard deviation (6 D), first difference (3 D), accelerometer window mean (3 D), and temperature features (3 D). Targets are 3-axis unsmoothed bias labels. Temporal models use overlapping sequences of length L, labeled at the last timestep.

MLP: [64, 64] hidden layers, ReLU, dropout 0.2. **LSTM**: 2-layer bidirectional, h=64, LayerNorm, residual projection, attention pooling. **TCN**: 5 weight-normalized dilated blocks [32, 64, 64, 128, 128], k=3, dilation [1, 2, 4, 8, 16] (RF=125 steps = 0.625 s), spatial dropout 0.2. **Transformer**: pre-norm, d=128, H=4, d_{ff}=512, 2 layers, `[CLS]` token, $\eta = 5 \times 10^{-4}$.

3.3. Training Protocol

Models are trained for 200 epochs with MSE loss and AdamW [37] (weight decay 10^{-4}). The training schedule uses cosine annealing (20-epoch warmup, $\eta_{\min}$=10^{-6}), gradient clipping ($\|g\|$=1.0), early stopping (patience 15), and mixed precision (AMP). The unified baseline uses RTX 3090 and RTX 4080 (PyTorch 2.5.1+cu121).

Each LOSO fold splits training data temporally (80/20, no shuffling). Features are z-normalized on training data only. Temporal-model sequences (stride 1) are constructed per source sequence to prevent cross-boundary leakage.

3.4. OOD Gating

MC Dropout [38] OOD gating (20 stochastic forward passes, 95th-percentile threshold) flags uncertain predictions and replaces them with zero correction. A 1-second grace period suppresses boundary artifacts. Full OOD results are reported in the Supplementary Material.

3.5. Physics-Informed Loss (PINN)

We test a physics-informed regularization term: given predicted bias $\hat{b}$, we integrate the corrected angular velocity $\omega_{\text{corr}} = \omega_{\text{meas}} - \hat{b}$ to obtain a predicted quaternion trajectory $\hat{q}$ and penalize the geodesic distance $\mathcal{L}_{\text{PINN}} = \frac{1}{N}\sum_k \arccos(|\hat{q}_k \cdot q_k^{\text{ref}}|)$ where q^{ref} is the ground-truth quaternion. The PINN loss is added to the primary data-fitting loss with weight $\lambda_{\text{phys}} \in \{10^{-3}, 10^{-2}, 10^{-1}, 1\}$. We evaluate PINN on EuRoC with LSTM and TCN. For each seed, the best λ is selected by *validation loss* within the LOSO training split (not on the held-out test sequence), and the corresponding test-set improvement is reported in Section 4.3.

3.6. Evaluation

We perform 5-fold LOSO on EuRoC and 6-fold LOSO on TUM-VI, each repeated with 17 random seeds on a single GPU per seed. The primary metric is attitude drift improvement:

$$\text{Imp}(\%) = \frac{\bar{e}_{\text{base}} - \bar{e}_{\text{corr}}}{\bar{e}_{\text{base}}} \times 100, \tag{1}$$

where $\bar{e}$ denotes mean final attitude error across folds. This drift-weighted metric avoids instability from near-zero baselines [39]. Statistical testing

Table 3: LOSO results (unified baseline, temporal splitting, unsmoothed labels, N=17 seeds).

Model	Dataset	Imp.%	Err. (°)	SD (°)	95% CI
TCN	EuRoC	96.19	5.70	0.29	[96.1, 96.3]
LSTM	EuRoC	95.80	6.29	0.49	[95.6, 96.0]
Trans.	EuRoC	95.63	6.53	0.71	[95.4, 95.9]
MLP	EuRoC	95.26	7.09	0.39	[95.1, 95.4]
LSTM	TUM-VI	21.57	116.83	10.01	[18.4, 24.8]
Trans.	TUM-VI	15.99	125.14	18.14	[10.2, 21.8]
TCN	TUM-VI	14.68	127.10	15.58	[9.7, 19.6]
MLP	TUM-VI	11.08	132.45	10.56	[7.7, 14.5]

uses paired permutation tests (10,000 permutations, Phipson–Smyth correction [29], Holm–Bonferroni family-wise control) and Cohen's d effect sizes. Cross-GPU consistency (<0.3 pp between RTX 3090 and 4080) is verified in Section 5.5. The full study comprises 2,811 unique runs (SHA-256 deduplicated): four architectures, two datasets, two feature representations, three losses, OOD gating, PINN regularization, iso-parameter and raw-input ablations (5–17 seeds per condition). Every configuration is repeated with 17 random seeds $\{7, 13, 42, 99, 123, 256, 314, 512, 777, 1234, 1999, 2024, 3141, 4242, 8888, 9999, 31415\}$.

4. Results

We report results from 272 unified baseline experiments (4 models × 2 datasets × 2 feature sets × 17 seeds) using unsmoothed labels. Supplementary material provides loss function, OOD gating, and Transformer ablation details.

4.1. Model Comparison

Table 3 and Fig. 1 present the primary LOSO results. On EuRoC, TCN leads (96.19%) followed by LSTM (95.80%), Transformer (95.63%), and MLP (95.26%). The Friedman test [40] (N=17) rejects equal model ranks (χ^2=32.29, p<0.001); five of six pairwise comparisons are significant after Holm–Bonferroni correction (Supplementary Table S1); only LSTM–Transformer fails to separate. Effect sizes range from $|d|$=0.82 (MLP–LSTM)

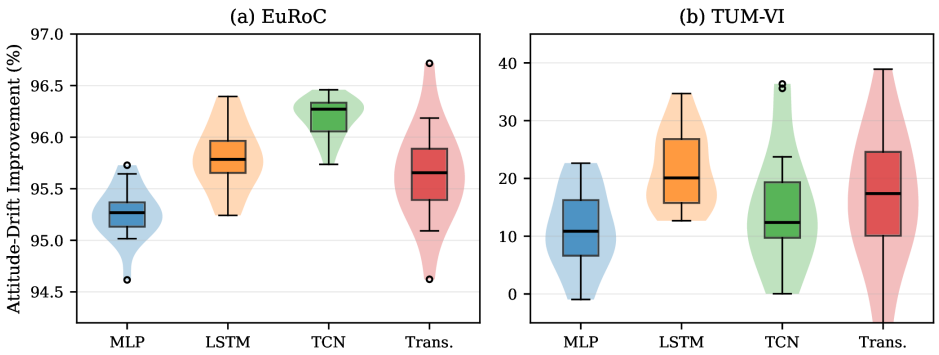

Figure 1: Attitude-drift improvement across N=17 seeds (single-GPU per seed, unsmoothed labels). (a) EuRoC: all models exceed 95%; TCN is the most accurate and the most consistent (smallest IQR). (b) TUM-VI: LSTM leads at 21.57%; all distributions are wider due to noisy pseudo-bias labels. Whiskers span the full seed range.

to $|d|$=4.09 (MLP–TCN). Despite statistical significance, the EuRoC spread is compressed to 0.93 pp, with all models exceeding 95%.

On TUM-VI, LSTM leads (21.57%) followed by Transformer (15.99%), TCN (14.68%), and MLP (11.08%); Friedman rejects (χ^2=12.04, p=0.007), but only LSTM–MLP reaches significance (p_{Holm}=0.004, d=1.52). Improvement is much lower (11–22% vs. 95–96%) because TUM-VI's pseudo-bias labels are inherently noisier (Section 3.1).

On EuRoC, the largest pairwise difference is MLP–TCN ($\Delta = -0.93$ pp, $d=-4.09$); on TUM-VI, only LSTM–MLP reaches significance ($\Delta=+10.49$ pp, d=1.52).

Across the two datasets, the contrast is consistent with dataset noise inflating inter-seed variance to the point where architecture differences become indistinguishable. This pattern is consistent with the variance decomposition in Section 5.1.

4.2. Room1 Pathology

All four models produce *negative* improvement on Room1 (−2,003% to −2,938%), inflating the 4.1° baseline to 87–126° across all 136 Room1 experiments. The root cause is that Room1's pseudo-bias labels are dominated by ground-truth differentiation noise: integrating MoCap-derived ω_{true} alone yields 52.9° error for Room1, *worse* than the raw gyroscope (4.1°), because quaternion differentiation amplifies high-frequency noise. Any learned correction adds noise to an already-accurate signal. When Room1 is excluded, all

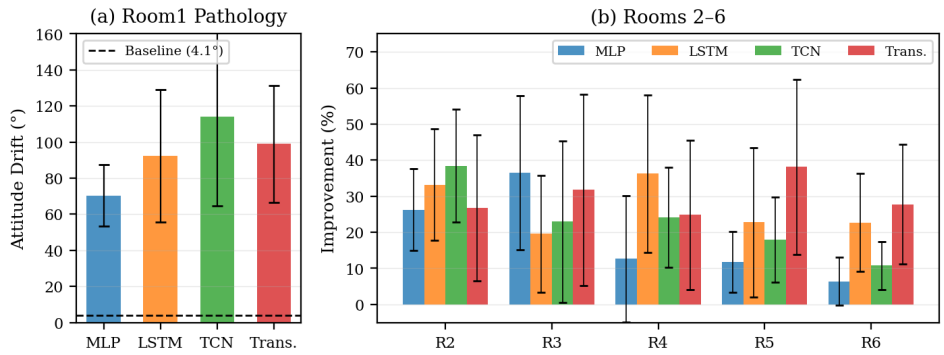

Figure 2: TUM-VI per-room breakdown (N=17 seeds). (a) Room1: all four models increase drift from the 4.1° baseline to >120° (negative improvement). (b) Rooms 2–6: LSTM leads on Rooms 3–5 (28–39%), the primary driver of its global TUM-VI advantage. Error bars: ±1 standard deviation across seeds.

Table 4: PINN quaternion regularization on EuRoC (unsmoothed labels, N=17). Base: seed-matched non-PINN MSE result; PINN: best λ per seed from $\{10^{-3}, 10^{-2}, 10^{-1}, 1\}$. Negative Δ indicates PINN degrades performance.

Model	N	Base (%)	PINN (%)	Δ (pp)	$\lvert d \rvert$	Sig.
LSTM	17	95.80	95.36	−0.43	0.64	**
TCN	17	96.19	95.32	−0.87	2.24	***

Paired t-test; $\lvert d \rvert$ = Cohen's d. TCN: 0/17 seeds improved by PINN.

models improve substantially (MLP 20.5%, LSTM 31.0%, TCN 28.4%, Transformer 28.0%), and the LSTM>TCN ranking is preserved. On Rooms 2–6, LSTM leads on 3 of 5 rooms (28–39%), driving its global TUM-VI advantage.

This pathology produces Simpson's paradox [41]: drift-weighted and per-fold-average metrics yield *opposite* rankings, which highlights the need for per-sequence reporting.

The Room1 result also indicates a practical boundary condition: when baseline drift is near the noise floor, *any* bias correction is harmful. OOD gating based on predictive variance [38] cannot detect this failure because it is a label-space anomaly (concept drift), not an input-space distributional shift. How to identify and flag such sequences before correction remains unresolved.

4.3. PINN Regularization

Physics-informed quaternion regularization degrades EuRoC performance for both architectures (Table 4). A confirmatory study includes 136 experiments with N=17 seeds and selects the best $\lambda \in \{10^{-3}, 10^{-2}, 10^{-1}, 1\}$ per seed. It shows LSTM -0.43 pp (p=0.008, $|d|$=0.64) and TCN -0.87 pp (p<0.001, $|d|$=2.24); TCN is degraded in all 17 seeds. A cross-GPU dose-response study (390 additional experiments) reveals a weakly negative Spearman correlation (ρ=-0.24, p=0.004, N=150 pooled conditions), ruling out under-regularization as the failure mode. This negative result is consistent with known gradient-level conflicts between data-fitting and physics terms [31, 30].

4.4. Loss Function

Comparing MSE, Huber (δ=0.005), and L1 losses, we find limited impact relative to architecture or dataset. On EuRoC, Huber marginally improves MLP (+0.60 pp), TCN (+0.39 pp), and Transformer (+1.42 pp) over MSE, while LSTM favors MSE. On TUM-VI, MSE dominates for MLP, LSTM, and Transformer; TCN is the sole exception, gaining +10.44 pp with Huber.

Overall, loss choice is less influential than dataset or architecture.

4.5. OOD Gating

MC-Dropout OOD gating with EMA-denoised labels improves three of four models on TUM-VI (+12–16 pp, d>1.2) but penalizes all models on EuRoC (-0.16 to -2.45 pp). With unsmoothed labels (136 runs, N=17), the TUM-VI benefit vanishes ($\Delta = -9.2$ to +5.4 pp), while EuRoC penalties widen (-0.3 to -5.3 pp, all p_{Holm}<0.001).

This preprocessing dependence suggests that OOD gating interacts with label noise: EMA denoising lowers validation variance and the resulting OOD threshold, inflating fallback frequency without improving detection accuracy.

5. Analysis

5.1. Variance Decomposition

A two-way ANOVA (model × dataset, N=17 seed means per cell) attributes $\eta^2 \approx 97\%$ of improvement variance to the dataset factor, with model architecture explaining <1% and the interaction explaining 0.2%. The model main effect (0.23%) and the interaction (0.20%) are both negligible compared to the dataset factor, meaning that architecture advantages do not meaningfully transfer across these two datasets. With only two dataset levels

this decomposition is descriptive rather than generalizable; it largely reflects the wide gap between clean-label EuRoC ($> 95\%$) and noisy-label TUM-VI ($< 22\%$).

Within our experimental scope, *label quality* dominates architecture choice [26], with a dataset-to-model variance ratio exceeding 400:1. Within our two-benchmark scope, switching from noisy pseudo-bias labels (TUM-VI) to batch-optimized labels (EuRoC) is associated with >70 pp improvement, while switching architectures on the same labels yields <1 pp. Within these benchmark conditions, ground-truth quality matters more than model selection.

5.2. Ranking Stability

We distinguish two levels of evidence: *seed-level robustness* (how stable is ranking across random initializations on the same data?) and *sequence-level generalization* (does the ranking hold across held-out sequences?). At the seed level (N=17), TCN holds EuRoC rank 1 in 12/17 seeds; LSTM holds TUM-VI rank 1 in 9/17. The Friedman tests (χ^2=32.29 and 12.04, p<0.001 and 0.007) treat seeds as repeated measures, which quantifies optimization stochasticity but should not be interpreted as 17 independent dataset-level replications. At the fold level (5 EuRoC sequences, 6 TUM-VI rooms), averaging across seeds within each fold, the per-fold model spread is 1.4–3.0 pp on EuRoC and 10–22 pp on TUM-VI Rooms 2–6 (Room1 excluded due to pathological 935 pp spread).

The fold-level sample size (N=5 or 6) is insufficient for formal hypothesis testing at α=0.05; the seed-level tests should therefore be read as evidence of *robust ranking under optimization noise*, not as confirmation of generalization to unseen IMU platforms.

5.3. Temporal Split and Label Preprocessing

Table 5 compares temporal (chronological) versus random train/validation splitting. On EuRoC, the effect is negligible for all models ($|\Delta| \leq 0.89$ pp, all $p > 0.4$). On TUM-VI, MLP gains a significant +13.58 pp (p_{Holm}=0.019) because random splitting leaks temporal structure into the validation set; MLP benefits most because it lacks temporal inductive bias and overfits to this leaked signal.

Label preprocessing also matters: removing EMA denoising improves all EuRoC models by +1.2–3.5 pp and *reverses* the TUM-VI ranking (MLP

Table 5: Effect of temporal splitting on LOSO performance (GPU-averaged, N=5, Holm correction over 8 tests).

Model	Dataset	Temp.	Rand.	Δ (pp)	p_{Holm}	Sig.
MLP	EuRoC	92.93	93.33	−0.40	0.577	n.s.
LSTM	EuRoC	94.18	94.47	−0.28	1.000	n.s.
TCN	EuRoC	95.23	95.49	−0.26	0.483	n.s.
Trans.	EuRoC	92.02	91.13	+0.89	0.589	n.s.
MLP	TUM-VI	30.48	16.90	+13.58	0.019	*
LSTM	TUM-VI	19.96	13.99	+5.97	1.000	n.s.
TCN	TUM-VI	19.61	16.87	+2.74	0.801	n.s.
Trans.	TUM-VI	11.90	10.22	+1.68	1.000	n.s.

Table 6: Label ablation: plain vs. GTSAM-smoothed labels (TUM-VI, GPU-averaged, N=5).

Model	Plain	GTSAM	Δ (pp)	p_{Holm}	Sig.
MLP	30.48	23.24	−7.23	0.401	n.s.
LSTM	19.96	13.34	−6.62	0.402	n.s.
TCN	19.61	7.96	−11.65	0.271	n.s.
Trans.	11.90	18.01	+6.11	0.358	n.s.

to LSTM at rank 1), because EMA acts as implicit regularization [42] for low-capacity models while degrading temporal models.

Table 6 evaluates GTSAM-smoothed versus EMA-denoised pseudo-bias labels on TUM-VI. Counter-intuitively, GTSAM smoothing hurts MLP (−7.23 pp), LSTM (−6.62 pp), and TCN (−11.65 pp), while marginally helping Transformer (+6.11 pp); none reach significance after Holm correction.

The unsmoothed labels used in our primary baseline avoid these preprocessing confounds. This finding cautions against treating label preprocessing as a neutral choice: a decision made before training can shift results by several percentage points and reverse the final ranking of architectures.

5.4. Within-Dataset Label Corruption

To isolate the effect of label quality from other benchmark-condition differences, we add synthetic Gaussian noise ($\sigma \in \{0, 0.01, 0.1, 0.5, 1.0\}$ rad/s) to the clean EuRoC bias labels and retrain LSTM and TCN (5 seeds each, 50 runs total). Performance degrades monotonically: at σ=0 both models

exceed 94%; at σ=0.1 (noise comparable to bias magnitude), improvement drops to ∼89%; at σ=1.0 (noise comparable to TUM-VI label noise), LSTM falls to 45.6% and TCN to 61.9%.

The architecture gap remains <4 pp at every noise level, while the label-noise effect spans >33 pp. Within this single dataset, the label-noise-to-architecture effect ratio is approximately 40:1, corroborating the cross-dataset variance decomposition (Section 5.1) with a controlled, within-dataset manipulation.

5.5. Cross-GPU Reproducibility

The unified baseline runs each seed on a single GPU (RTX 3090/4080, PyTorch 2.5.1; cross-GPU variance <0.3 pp). A separate study on RTX 5090 and H100 (390 PINN experiments) reveals condition-dependent reproducibility: LSTM EuRoC agrees well (ICC(2,1)=0.94), while Transformer TUM-VI is irreproducible (ICC=−0.13), likely due to floating-point non-associativity [43].

5.6. Iso-Parameter Comparison

To disentangle architecture from capacity, we train MLP, LSTM, and TCN at two matched scales (∼5K and ∼60K parameters) on EuRoC (30 runs). At 5K, the maximum spread is 0.32 pp; at 60K, 1.11 pp. LSTM *degrades* from 5K to 60K (−0.48 pp), suggesting overfitting. When parameter count is equalized, architecture differences shrink to <1 pp, confirming that the ∼0.9 pp spread in Table 3 is dominated by capacity rather than architecture.

Supplementary material reports two additional analyses: (i) a closed-loop hybrid ML+ESKF evaluation (100 evaluations) that preserves the TCN>LSTM ordering on EuRoC; and (ii) a raw-input ablation (7-dim raw IMU, 40 experiments) confirming that engineered features do not bias rankings. All three analyses support the conclusion that dataset quality, not model architecture, is the dominant factor.

6. Discussion

Benchmark condition outweighs architecture choice.. The cross-dataset variance decomposition ($\eta^2 \approx 97\%$ for dataset) dwarfs model and interaction effects. A within-dataset label corruption experiment (Section 5.4) confirms that this is largely driven by label quality: degrading EuRoC labels to TUM-VI-like noise levels reduces performance by 33–49 pp, while switching architectures at any fixed noise level changes results by <4 pp (ratio ≈40:1).

Rankings are also sensitive to preprocessing: removing EMA denoising improves all EuRoC models by +1.2–3.5 pp while *reversing* the TUM-VI ranking, because EMA acts as implicit regularization [42] for low-capacity models while degrading temporal models that exploit high-frequency label structure.

Physics-informed regularization underperforms.. The quaternion PINN penalty consistently degrades both LSTM and TCN (Table 4), with a negative dose-response consistent with gradient-level conflicts between data-fitting and physics terms [31, 32].

Lie-group dynamics [21] may fare better by avoiding the soft-penalty formulation entirely.

Room1 pathology and Simpson's paradox.. Room1's 4.1° baseline drift is near the noise floor; integrating MoCap-derived ω_{true} alone yields 52.9° error (*worse* than the raw gyroscope) because quaternion differentiation amplifies noise. Any learned correction adds noise to an already accurate signal. MC-Dropout OOD gating [38] cannot detect this anomaly because it resides in the label space, not the input space where epistemic uncertainty operates [44].

This behavior produces Simpson's paradox [41]: drift-weighted and per-fold-average metrics yield opposite rankings.

Limitations.. (1) Only two datasets; generalization to other IMU platforms is untested. (2) Closed-loop ESKF evaluation confirms rankings but full VIO integration remains untested. (3) Iso-parameter matching (Supplementary) shows architecture differences shrink to <1 pp when capacity is equalized, but confounds receptive field. (4) MC Dropout OOD gating misses label-space anomalies; a preliminary heteroscedastic (beta-NLL) loss experiment [45] showed training instability on EuRoC and no benefit on TUM-VI. (5) Mamba, diffusion models [20], and foundation-model pre-training are not evaluated.

Uncontrolled confounds shifted results by up to 51 pp in our early experiments [26]; future benchmarks should fix seeds, enforce temporal splitting, and control label preprocessing. We encourage reporting experiment counts, random seeds, and GPU specifications alongside performance metrics.

7. Conclusion

We compared MLP, LSTM, TCN, and Transformer architectures for MEMS gyroscope bias estimation in 2,811 experiments across two public datasets (EuRoC, TUM-VI) and 17 random seeds. Under controlled confounds,

TCN achieves the best attitude-drift improvement on EuRoC (96.19%), while LSTM leads on TUM-VI (21.57%) with little statistically significant separation between models.

Variance decomposition shows that benchmark condition explains ≈97% of performance variation; model architecture accounts for less than 1%. A within-dataset label corruption ablation confirms that label quality is the primary driver (noise-to-architecture effect ratio ≈40:1). Physics-informed SO(3) regularization fails to improve estimation quality in 136 confirmatory experiments, and label preprocessing choices can shift results by 1–3 pp and reverse model rankings.

Four recommendations follow: (1) invest in label quality before model complexity (dataset-to-model variance ratio exceeds 400:1); (2) use temporal splitting to avoid leaking temporal structure (+13.58 pp for MLP on TUM-VI); (3) develop label-space anomaly detectors for near-zero-drift sequences, as MC-Dropout OOD gating proved insufficient; (4) control label preprocessing (e.g., EMA denoising), which can reverse model rankings.

Future work should expand to larger, more diverse IMU corpora [5, 27] and validate within full visual-inertial odometry pipelines. Code and experiment logs are available at `https://github.com/thc1006/mems_ai_calibration_pytorch`.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used Claude (Anthropic) to assist with data-analysis automation, experiment orchestration, and limited language editing. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

[1] O. J. Woodman, An introduction to inertial navigation, Tech. Rep. UCAM-CL-TR-696, University of Cambridge, Computer Laboratory (2007). doi:10.48456/tr-696.

[2] N. El-Sheimy, A. Youssef, Inertial sensors technologies for navigation applications: State of the art and future trends, Satellite Navigation 1 (1) (2020) 1–21. doi:10.1186/s43020-019-0001-5.

[3] Y. Long, Z. Liu, C. Hao, F. Ayazi, MEMS gyroscope multi-feature calibration using machine learning technique, arXiv preprint arXiv:2410.07519 (2024).

[4] C. Cui, J. Zhao, H. Long, R. Zhang, TinyGC-Net: An extremely tiny network for calibrating MEMS gyroscopes, arXiv preprint arXiv:2403.02618 (2024).

[5] S. Zhao, S. Zhou, R. Blanchard, Y. Qiu, W. Wang, S. Scherer, Tartan IMU: A light foundation model for inertial positioning in robotics, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 22520–22529.

[6] N. Cohen, I. Klein, Inertial navigation meets deep learning: A survey of current trends and future directions, Results in Engineering 24 (2024) 103565. doi:10.1016/j.rineng.2024.103565.

[7] R. Buchanan, V. Agrawal, M. Camurri, F. Dellaert, M. Fallon, Deep IMU bias inference for robust visual-inertial odometry with factor graphs, IEEE Robotics and Automation Letters 8 (1) (2023) 41–48. doi:10.1109/LRA.2022.3222956.

[8] C. Bergmeir, J. M. Benítez, On the use of cross-validation for time series predictor evaluation, Information Sciences 191 (2012) 192–213. doi:10.1016/j.ins.2011.12.028.

[9] V. Cerqueira, L. Torgo, I. Mozetič, Evaluating time series forecasting models: An empirical study on performance estimation methods, Machine Learning 109 (11) (2020) 1997–2028. doi:10.1007/s10994-020-05910-7.

[10] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9 (8) (1997) 1735–1780. doi:10.1162/neco.1997.9.8.1735.

[11] S. Bai, J. Z. Kolter, V. Koltun, An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, arXiv preprint arXiv:1803.01271 (2018).

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, Vol. 30, 2017, pp. 5998–6008.

[13] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

[14] J. L. Ba, J. R. Kiros, G. E. Hinton, Layer normalization, arXiv preprint arXiv:1607.06450 (2016).

[15] M. Brossard, S. Bonnabel, A. Barrau, Denoising IMU gyroscopes with deep learning for open-loop attitude estimation, IEEE Robotics and Automation Letters 5 (3) (2020) 4796–4803. doi:10.1109/LRA.2020.3003256.

[16] W. Liu, D. Caruso, E. Ilg, J. Dong, A. I. Mourikis, K. Daniilidis, V. Kumar, J. Engel, TLIO: Tight learned inertial odometry, IEEE Robotics and Automation Letters 5 (4) (2020) 5653–5660. doi:10.1109/LRA.2020.3007421.

[17] Y. Qiu, C. Wang, C. Xu, Y. Chen, X. Zhou, Y. Xia, S. Scherer, AirIMU: Learning uncertainty propagation for inertial odometry, arXiv preprint arXiv:2310.04874 (2023).

[18] F. Huang, Z. Wang, L. Xing, C. Gao, A MEMS IMU gyroscope calibration method based on deep learning, IEEE Transactions on Instrumentation and Measurement 71 (2022) 1–9. doi:10.1109/TIM.2022.3160538.

[19] R. Li, C. Fu, W. Yi, X. Yi, Calib-Net: Calibrating the low-cost IMU via deep convolutional neural network, Frontiers in Robotics and AI 8 (2022) 772583. doi:10.3389/frobt.2021.772583.

[20] S. Zhou, S. Katragadda, G. Huang, Learning IMU bias with diffusion model, in: IEEE International Conference on Robotics and Automation (ICRA), 2025.

[21] B. Liu, T.-Y. Lin, W. Zhang, M. Ghaffari, Debiasing 6-DOF IMU via hierarchical learning of continuous bias dynamics, in: Robotics: Science and Systems (RSS), 2025.

[22] W. Ye, T. Liu, Z. Jiang, M. Wu, A robust transformer-based error compensation method for gyroscope of IMUs, Journal of Field Robotics (2026). doi:10.1002/rob.70082.

[23] Y. Liu, W. Liang, J. Cui, LGC-Net: Lightweight gyroscope errors compensation network for effective attitude estimation, in: 42nd Chinese Control Conference (CCC), 2023, pp. 4041–4046. doi:10.23919/CCC58697.2023.10241087.

[24] R. J. Hyndman, G. Athanasopoulos, Forecasting: Principles and Practice, 3rd Edition, OTexts, Melbourne, Australia, 2021.

[25] J. Demšar, Statistical comparisons of classifiers over multiple data sets, Journal of Machine Learning Research 7 (2006) 1–30.

[26] X. Bouthillier, P. Delaunay, M. Bronzi, A. Trofimov, B. Nichyporuk, J. Szeto, N. Sepahvand, E. Raff, K. Madan, V. Voleti, S. E. Kahou, V. Michalski, T. Arbel, C. Pal, G. Varoquaux, P. Vincent, Accounting for variance in machine learning benchmarks, in: Proceedings of Machine Learning and Systems (MLSys), Vol. 3, 2021, pp. 747–769.

[27] P. Ferreira, P. Costa, IF-D: A high-frequency, general-purpose inertial foundation dataset for self-supervised learning, arXiv preprint arXiv:2510.09539 (2025).

[28] P. Bayle, A. Bayle, L. Janson, L. Mackey, Cross-validation confidence intervals for test error, in: Advances in Neural Information Processing Systems, Vol. 33, 2020, pp. 16339–16350.

[29] B. Phipson, G. K. Smyth, Permutation P-values should never be zero: Calculating exact P-values when permutations are randomly drawn, Statistical Applications in Genetics and Molecular Biology 9 (1) (2010). doi:10.2202/1544-6115.1585.

[30] S. Wang, X. Yu, P. Perdikaris, When and why PINNs fail to train: A neural tangent kernel perspective, Journal of Computational Physics 449 (2022) 110768. doi:10.1016/j.jcp.2021.110768.

[31] A. S. Krishnapriyan, A. Gholami, S. Zhe, R. M. Kirby, M. W. Mahoney, Characterizing possible failure modes in physics-informed neural networks, in: Advances in Neural Information Processing Systems, Vol. 34, 2021, pp. 26548–26560.

[32] C. Xu, D. Liu, A. Nassereldine, J. Xiong, FP64 is all you need: Rethinking failure modes in physics-informed neural networks, in: Advances in Neural Information Processing Systems (NeurIPS), 2025.

[33] Q. Liu, M. Chu, N. Thuerey, ConFIG: Towards conflict-free training of physics informed neural networks, in: International Conference on Learning Representations (ICLR), 2025.

[34] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, R. Siegwart, The EuRoC micro aerial vehicle datasets, The International Journal of Robotics Research 35 (10) (2016) 1157–1163. doi:10.1177/0278364915620033.

[35] F. Dellaert, Factor graphs and GTSAM: A hands-on introduction, Tech. rep., Georgia Institute of Technology (2012).

[36] D. Schubert, T. Goll, N. Demmel, V. Usenko, J. Stückler, D. Cremers, The TUM VI benchmark for evaluating visual-inertial odometry, in: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2018, pp. 1680–1687. doi:10.1109/IROS.2018.8593419.

[37] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations (ICLR), 2019.

[38] Y. Gal, Z. Ghahramani, Dropout as a Bayesian approximation: Representing model uncertainty in deep learning, in: International Conference on Machine Learning (ICML), 2016, pp. 1050–1059.

[39] R. J. Hyndman, A. B. Koehler, Another look at measures of forecast accuracy, International Journal of Forecasting 22 (4) (2006) 679–688. doi:10.1016/j.ijforecast.2006.03.001.

[40] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, Journal of the American Statistical Association 32 (200) (1937) 675–701. doi:10.1080/01621459.1937.10503522.

[41] O. Salaudeen, L. Zhang, K. Alhamoud, S. Beery, M. Ghassemi, Aggregation hides out-of-distribution generalization failures from spurious correlations, in: Advances in Neural Information Processing Systems, Vol. 38, 2025.

[42] L. Semenova, H. Chen, R. Parr, C. Rudin, A path to simpler models starts with noise, in: Advances in Neural Information Processing Systems, Vol. 36, 2023.

[43] S. Shanmugavelu, M. Taillefumier, C. Culver, O. Hernandez, M. A. Coletti, A. Sedova, Impacts of floating-point non-associativity on reproducibility for HPC and deep learning applications, in: SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2024, pp. 170–179. doi:10.1109/SCW63240.2024.00028.

[44] A. Djupskås, A. J. Stasik, S. Riemer-Sørensen, Unreliable uncertainty estimates with Monte Carlo dropout, in: Northern Lights Deep Learning Conference (NLDL), 2026.

[45] M. Seitzer, A. Tavakoli, D. Antic, G. Martius, On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks, in: International Conference on Learning Representations, 2022.
URL https://openreview.net/forum?id=aPOpXlnV1T