Predicting Cardiovascular Drift from 3D Running Pose via Spatio-Temporal Graph Attention

Yuval David[1] **Barak Gahtan**[1] **Alex M. Bronstein**[1,2]
[1]Technion – Israel Institute of Technology
[2]ISTA – Institute of Science and Technology Austria
{yuval.david, barakgahtan, bron}@cs.technion.ac.il

Abstract

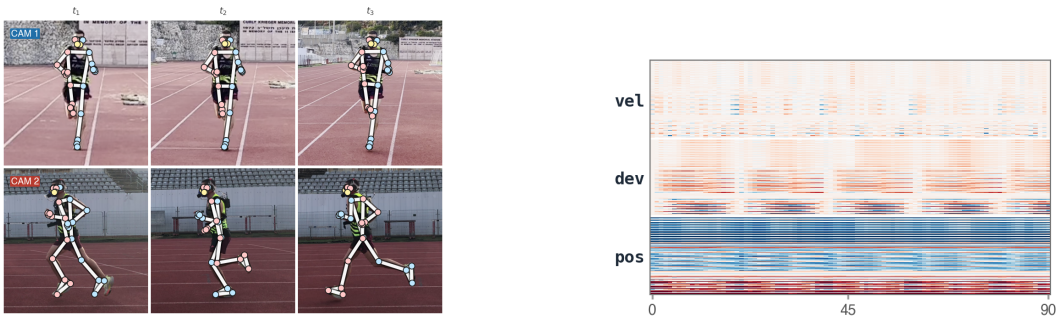
Cardiovascular drift, the progressive rise in heart rate during sustained exercise at stable intensity is a sensitive indicator of physiological strain that traditionally requires continuous chest-strap instrumentation. We show that this latent quantity can be estimated from video alone, with no physiological sensor required at inference time. We propose **AsymSGAT**, a compact ($\sim$24 K parameter) spatio-temporal graph attention network that predicts cumulative cardiac drift from 3D running pose under a *session-calibrated* protocol: one baseline lap of video establishes a personalised pose reference, after which drift on all subsequent laps is predicted from kinematics alone; heart-rate measurements inform training supervision but are absent at deployment. Under 13-fold leave-one-subject-out evaluation (n=39 lap-level samples), AsymSGAT achieves **0.62 bpm** MAE and Pearson r=0.95, a 77% reduction over the predict-mean baseline (p<0.001) and a factor of $\geq$3.7$\times$ improvement over all neural and classical alternatives. Three independent controls rule out subject memorisation: replacing all inputs with Gaussian noise collapses MAE to 6.38 bpm; an identity probe recovers subject identity from the learned representation at only 17.9% versus 69.2% from raw pose; and LOSO structurally prevents held-out subjects from appearing in training. Face-based remote photoplethysmography is inapplicable here due to facial occlusion throughout the recordings. These results establish that video-derived running kinematics carry actionable information about latent cardiovascular state, opening a direction toward non-contact within-session exercise monitoring.

1 Introduction

Cardiovascular drift (CV drift), the progressive rise in heart rate during sustained submaximal exercise at nominally stable intensity, is a hallmark of accumulating physiological strain, driven by plasma-volume reduction, thermoregulatory load, and declining stroke volume [Coyle and Gonzalez-Alonso, 2001]. As a sensitive early indicator of overexertion and heat stress that precedes performance collapse, exertional heat illness, and, in extreme cases, cardiac events [Périard et al., 2021], real-time CV drift estimation has clear value for training-load management, fatigue-aware pacing, and safety monitoring. Yet current approaches require direct physiological instrumentation.

Wearable sensors (chest straps, wrist-worn PPG) yield accurate, continuous measurements but impose a compliance burden, limit scalability, and are impractical in settings where instrumentation is infeasible or undesirable [Reiss et al., 2019, Nazaret et al., 2023, Gahtan et al., 2025]. Remote photoplethysmography (rPPG) offers a non-contact alternative by recovering instantaneous heart rate from subtle facial skin-colour fluctuations [Chen and McDuff, 2018, Liu et al., 2020], but requires an unobstructed, near-stationary face at sufficient resolution, conditions that fail systematically under the vigorous whole-body motion, variable illumination, and facial occlusion characteristic of outdoor

Preprint.



(a) Video with 3D pose overlay

(b) 9-channel input tensor

Figure 1: **Inputs to AsymSGAT.** For each camera view, a short running clip is converted into a 33-joint 3D pose sequence (a) and then into a $9 \times T \times J$ tensor encoding root-centered position, deviation from baseline, and velocity, with T=90 frames and J=33 joints (b).

running. To our knowledge, no prior feasibility study has demonstrated estimation of cardiovascular state during vigorous full-body exercise without directly instrumenting the athlete.

We propose to close this gap via a different signal modality: whole-body kinematics. Exercise-induced fatigue produces well-documented biomechanical adaptations (altered arm swing, reduced vertical stiffness, changing ground-contact mechanics, and progressive bilateral asymmetry [Clermont et al., 2019, Zandbergen et al., 2023, Zifchock et al., 2008]) that share a common physiological origin with CV drift. Advances in markerless pose estimation now make it possible to extract stable 3D skeleton sequences from ordinary monocular video in real time [Bazarevsky et al., 2020]. Together, these observations motivate a new sensing paradigm: *inferring an internal cardiovascular trajectory from externally observable body movement*, without attaching any device to the athlete.

Approach. Our method operates under a *session-calibrated* protocol. One baseline lap of video per session is used to compute a personalised pose reference; drift on all subsequent laps is then predicted from kinematics alone using two fixed cameras. Ground-truth drift labels are derived from a chest-strap monitor used only during data collection; no physiological sensor is needed at inference. Lap index and video-derived speed are the only auxiliary inputs.

Contributions. **(1) Task formulation.** We introduce session-calibrated cumulative cardiac-drift estimation from dual-view running video as a continuous regression problem under controlled track conditions, requiring only a brief baseline observation per session at inference time. **(2) Architecture.** We propose **AsymSGAT**, a compact ($\sim$24 K parameter) spatio-temporal graph attention network whose spatial backbone uses dynamic multi-head graph attention [Brody et al., 2022] over the anatomical skeleton, learning per-sample edge weights. The model further incorporates FiLM conditioning on session-level context, temporal attention pooling, and an explicit bilateral-asymmetry descriptor. A shared-weight backbone processes both camera views; the two scalar predictions are averaged to produce a single per-lap estimate, exploiting complementary geometric perspectives without doubling the parameter count. **(3) Empirical validation.** Under leave-one-subject-out evaluation on 13 runners (n=39 lap-level samples), AsymSGAT achieves **0.62 bpm** MAE (r=0.95, -77% vs. predict-mean, $p<0.001$), outperforming all baselines by a factor of $\geq 3.7\times$. Three independent controls rule out subject memorisation: a noise-input collapse (MAE collapses to 6.38 bpm), a penultimate-layer identity probe, and the structural LOSO guarantee (Section 4.6).

2 Related Work

Heart-rate estimation: from sensors to video. Heart-rate prediction spans a spectrum of sensing modalities. Wrist-worn PPG with motion-artifact suppression achieves 1.5–2.5 bpm MAE during controlled exercise [Reiss et al., 2019]; personalised wearable models further improve HR response prediction [Nazaret et al., 2023]; and recent work coupling physiologically constrained ODEs with neural Kalman filtering reaches 2.81 bpm at 1 s resolution [Gahtan et al., 2025]. At the non-contact end, remote photoplethysmography (rPPG) recovers instantaneous heart rate from facial skin-colour

fluctuations [Chen and McDuff, 2018, Liu et al., 2020, Wang et al., 2017, de Haan and Jeanne, 2013]. Outdoor distance running compounds these difficulties: vigorous whole-body motion induces motion blur and illumination variation; camera-to-subject distances at typical field positions reduce the face to a sub-critical pixel footprint; and facial occlusion is common in outdoor athletic settings. Applied to our footage, POS [Wang et al., 2017] and CHROM [de Haan and Jeanne, 2013] yield 102–106 bpm MAE, confirming that face-based sensing is not viable in this regime. This motivates estimating a *cumulative cardiovascular quantity* from *full-body 3D pose*, which is accessible regardless of facial visibility.

Video-based physiology and skeleton modelling. A growing body of work demonstrates that observable body motion encodes latent internal states. Peng et al. [2022] regress instantaneous caloric expenditure from video; Schneider et al. [2024] learn a mapping from pose to OpenSim-derived muscle activations; gait-video models estimate clinical scalars such as cerebellar ataxia severity [Eguchi et al., 2025]. We extend this programme from the musculoskeletal to the cardiovascular domain, and from instantaneous predictions to *cumulative* quantities that require integrating progressive, subtle kinematic changes over minutes of exercise. On the architectural side, spatio-temporal graph convolutional networks provide a natural inductive bias for skeleton-based action understanding [Yan et al., 2018], with subsequent advances in adaptive topology [Chen et al., 2021] and multi-stream fusion [Shi et al., 2019]. Graph Attention Networks [Veličković et al., 2018] replace fixed normalised adjacency with learned per-edge attention weights; GATv2 [Brody et al., 2022] addresses a theoretical expressiveness limitation of the original formulation by applying the non-linearity before the attention projection, making the attention function strictly more expressive. Feature-wise Linear Modulation (FiLM) conditions graph and visual networks on auxiliary context [Brockschmidt, 2020, Perez et al., 2018]. Action recognition is a coarse categorisation task in which sequences from the same class are near-identical; physiological regression demands the opposite sensitivity — detecting small within-class kinematic variations while being robust to large between-subject differences. This motivates our architectural choices: an anatomically structured graph topology with dynamic graph attention over edges, FiLM conditioning on session-level baseline pose, and an explicit bilateral-asymmetry branch.

Running fatigue and cardiac drift. Sustained running induces systematic biomechanical adaptations: altered stride dynamics, reduced vertical stiffness, increased ground-contact time, and progressive bilateral asymmetry [Clermont et al., 2019, Zandbergen et al., 2023, Zifchock et al., 2008]. Wearable-IMU studies have established categorical fatigue-classification baselines [Strohrmann et al., 2012, Marotta et al., 2021, Chang et al., 2023, Dimmick et al., 2023]. The only prior machine-learning work targeting cardiovascular drift directly is Barsumyan et al. [2025], which applies a Variational Gaussian Process to binary classification (responder vs. non-responder) from sensor-derived heart rate; all existing drift-correction models assume a contact HR sensor at inference [Papini et al., 2024, Nazaret et al., 2023]. The present work is, to our knowledge, the first feasibility study demonstrating that continuous cumulative cardiac drift can be estimated from running video without any physiological sensor at inference.

Small-sample methodology in biomedical ML. Record-wise cross-validation inflates accuracy when multiple observations per subject are pooled [Saeb et al., 2017]; subject-wise (LOSO) evaluation is the methodologically sound alternative. Best-practice guidance [Schaefer et al., 2020] codifies subject-blocked evaluation and compact models with strong domain priors. Our design follows these standards: 13-fold LOSO, subject-blocked bootstrap CIs, permutation test, and Wilcoxon tests on per-subject MAEs.

3 Method

We formulate cardiac drift estimation as a regression problem over short pose clips and address it with a compact spatio-temporal graph attention network that encodes physiological priors. This section describes the prediction target, dataset, input representation, and model architecture.

Prediction target. For lap $i \geq 2$ in a running session, cumulative cardiac drift is defined as

$$y_i = \Delta\mathrm{HR}(i) = \mathrm{HR}(i) - \mathrm{HR}(1), \tag{1}$$

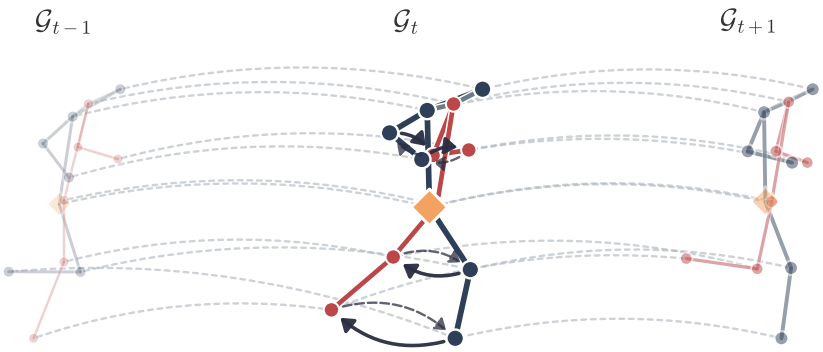


Figure 2: Spatio-temporal pose graph. Consecutive frames are linked by temporal edges (dashed grey). Left-side joints (blue) and right-side joints (red) share standard anatomical edges; the focus frame ($\mathcal{G}_t$) additionally shows the four **cross-body bilateral edges** as directed arrows. The root joint (left hip, $\diamondsuit$, orange) anchors the skeleton.

where $\mathrm{HR}(i)$ is the mean heart rate during lap i from a chest-strap sensor used only at training time. No physiological measurement is required at inference. Targets range up to 16 bpm (mean 7.4, std 3.8).

Data collection. Thirteen adult runners completed a repeated-lap protocol on a flat outdoor 400 m track at self-selected pace. Each session was recorded simultaneously from two stationary cameras positioned at distinct points on the circuit, at 30 fps. Heart rate was recorded continuously (Garmin HRM-Pro, 1 Hz); lap-average HR is computed over $\sim$120 samples per lap. Participants wore metabolic masks, resulting in partial facial occlusion. Monocular 3D pose was extracted from each camera with BlazePose (33 keypoints) [Bazarevsky et al., 2020] with hip-relative root subtraction. Each runner completed four laps; lap 1 defines the within-session baseline. The dataset comprises n=39 samples: 13 subjects $\times$ 3 laps (laps 2–4), each with two camera views.

Input representation. Each pose clip spans T=90 frames ($\approx$3 s at 30 fps) with J=33 joints, root-centred at the hip. We construct a 9-channel spatio-temporal tensor $\mathbf{X}_i \in \mathbb{R}^{9\times T\times J}$ by concatenating three feature groups: root-centred 3D coordinates (channels 1–3), frame-wise deviation from the session's baseline mean pose (channels 4–6), and frame-to-frame velocity (channels 7–9). Channels 4–6 require the first-lap pose from the same session, instantiating the session-calibrated protocol. A scalar conditioning vector $\mathbf{a}_i = [\ell_i, v_i, v_i^{\text{base}}] \in \mathbb{R}^3$ encodes normalised lap index and video-derived speed estimates; this vector is injected exclusively via FiLM modulation described below. Two such tensors $(\mathbf{X}_i^{(1)}, \mathbf{X}_i^{(2)})$ are constructed, one per camera view, and provided jointly as input.

Spatial graph structure. The 33-joint skeleton is represented as a graph $\mathcal{G} = (V, E)$ using the anatomical connectivity of the BlazePose skeleton, augmented with four cross-body edges between contralateral joint pairs (elbows, wrists, knees, ankles). These cross-body connections provide a structural prior for bilateral coordination, encouraging the network to compare left- and right-side activations during spatial message passing (Figure 2). The union of all edges, including self-loops, defines the attention topology: only connected joints exchange information.

Graph attention backbone. AsymSGAT processes $\mathbf{X}_i$ through five spatio-temporal blocks of channel widths $[8, 16, 16, 24, 24]$ with four attention heads, totalling 23,912 parameters. The spatial aggregation in each block is a GATv2 graph attention layer [Brody et al., 2022]: for each directed edge $(j \to i)$ in the graph, the attention coefficient is

$$e_{ij}^h = \left(\mathbf{a}^h\right)^\top \mathrm{LeakyReLU}\left(\mathbf{W}_l^h\,\mathbf{h}_i + \mathbf{W}_r^h\,\mathbf{h}_j\right), \tag{2}$$

where $\mathbf{a}^h$, $\mathbf{W}_l^h$, $\mathbf{W}_r^h$ are per-head learnable parameters. Attention weights are softmax-normalised over each node's neighbourhood; edges outside the anatomical topology receive $-\infty$ so the edge set is anatomically constrained. This is strictly more expressive than the original GAT [Veličković

et al., 2018]. Following spatial attention, each block applies temporal convolution (kernel 9), batch normalisation, ReLU, dropout, and a residual connection; stride 2 at blocks 3 and 5 reduces the sequence from 90 to 23 frames.

FiLM conditioning. Feature-wise Linear Modulation [Perez et al., 2018] is applied at the two strided blocks:

$$\mathrm{FiLM}(\mathbf{h}, \mathbf{a}_i) = \mathbf{h} \odot \big(1 + \gamma(\mathbf{a}_i)\big) + \beta(\mathbf{a}_i), \tag{3}$$

where γ and β are zero-initialised linear projections. This injects session-level context into intermediate activations while preserving the pose stream as the primary information pathway.

Temporal attention pooling. The temporal dimension is aggregated via learned attention weights,

$$\alpha_t = \frac{\exp(\mathbf{w}^\top \tanh(\mathbf{W}\mathbf{h}_t))}{\sum_\tau \exp(\mathbf{w}^\top \tanh(\mathbf{W}\mathbf{h}_\tau))}, \quad \mathbf{z}_{\mathrm{att}} = \textstyle\sum_t \alpha_t \mathbf{h}_t, \tag{4}$$

allowing the model to weight gait-cycle phases selectively rather than pooling uniformly.

Bilateral asymmetry descriptor. Fatigue-induced increases in bilateral asymmetry [Zifchock et al., 2008] motivate an explicit descriptor $\mathbf{z}_{\mathrm{asym}} = \bar{\mathbf{h}}_L - \bar{\mathbf{h}}_R \in \mathbb{R}^{24}$, computed from mean left- and right-side joint activations after the final block. This complements the cross-body graph edges: those edges provide structural pathways for bilateral information exchange during attention, while $\mathbf{z}_{\mathrm{asym}}$ exposes the resulting left-right divergence directly to the regression head. The regressor operates on $\mathbf{z} = [\mathbf{z}_{\mathrm{att}}; \mathbf{z}_{\mathrm{asym}}; \mathbf{a}_i] \in \mathbb{R}^{51}$ via $\mathrm{FC}(51{\to}24) \to \mathrm{ReLU} \to \mathrm{FC}(24{\to}1)$, with a zero-initialised auxiliary shortcut $\mathbf{w}_{\mathrm{aux}}^\top \mathbf{a}_i$ added to the scalar output (see Appendix D for a schematic).

Dual-view fusion. The same set of weights processes both camera views. Given two pose sequences $\mathbf{X}_i^{(1)}$ and $\mathbf{X}_i^{(2)}$ from distinct viewpoints, the final prediction is

$$\hat{y}_i = \tfrac{1}{2}\big[f_\theta\Big(\mathbf{X}_i^{(1)}, \mathbf{a}_i\Big) + f_\theta\Big(\mathbf{X}_i^{(2)}, \mathbf{a}_i\Big)\big], \tag{5}$$

where f_θ denotes the full AsymSGAT forward pass. Weight sharing enforces view-agnostic representations while exploiting the complementary geometric perspectives of the two cameras at inference. The total parameter count remains 23,912, since no additional weights are introduced by the dual-view fusion.

Training. Models are trained with Huber loss (δ=1.5) and AdamW (lr=$2{\times}10^{-3}$, weight decay $2{\times}10^{-3}$) with cosine annealing over 300 epochs, batch size 4, and early stopping (patience 40). Each batch consists of dual-view pose pairs; both views are augmented independently. Dropout rates are 0.15 in the graph blocks and 0.5 in the regression head. Training augmentations comprise temporal crop (80–100%), additive Gaussian pose noise (σ=0.01), left-right skeletal flip, and scale jitter (0.95–$1.05\times$). Per-fold predictions are the mean across 5 independent random seeds.

4 Experiments

4.1 Evaluation Protocol

All experiments use leave-one-subject-out (LOSO) cross-validation across all 13 subjects, with input features z-normalised per channel using training-fold statistics. MAE and Pearson r are reported on pooled held-out predictions with subject-blocked bootstrap 95% confidence intervals (subjects resampled, not laps) and per-fold dispersion statistics. Statistical significance is assessed via a subject-level permutation test (10,000 shuffles) and Wilcoxon signed-rank tests on per-subject MAEs. Following best practice for small-sample evaluation [Schaefer et al., 2020], each fold is trained with 5 independent random seeds and per-fold predictions are averaged across seeds to reduce initialisation variance. Appendix C.3 confirms that the per-seed variation is substantially smaller than the across-subject variation, so subject heterogeneity — not initialisation sensitivity — is the dominant source of uncertainty in the reported results.

Table 1: Method comparison under 13-fold LOSO (n=39). ΔMAE is relative to predict-mean. Subject-blocked bootstrap 95% CIs are reported for AsymSGAT; the CI upper bound (0.96 bpm) lies below all baseline point estimates. Wilcoxon signed-rank tests on per-subject MAEs confirm significant advantage over Ridge (W=0, p=1.2×10^{-4}) and predict-mean (W=0, p=1.2×10^{-4}).

Method	Input	MAE	r	ΔMAE
AsymSGAT	pose×2+aux	**0.62** [0.33, 0.96]	**0.95** [0.92, 0.98]	−77%
Ridge	biomech+aux	2.30	0.53	−13%
RF	biomech+aux	2.33	0.49	−12%
Per-lap-mean	—	2.35	0.44	−11%
Linear	aux only	2.55	0.22	−4%
Predict-mean	—	2.65	—	—

Dataset structure. The dataset comprises n=39 samples: 13 subjects × 3 laps (laps 2–4). Each sample corresponds to one lap and provides two synchronised camera views from distinct circuit positions as input. LOSO is performed at the subject level, withholding both camera views of all laps of a held-out subject from training jointly.

4.2 Baselines

We compare against classical baselines of increasing complexity:

- **Predict-mean:** per-fold constant prediction equal to the training-set mean drift.
- **Linear (metadata):** linear regression on auxiliary features only (no pose).
- **Ridge:** ridge regression on 20 hand-crafted biomechanical descriptors averaged across both views, concatenated with auxiliary features.
- **Random Forest:** 100-tree ensemble (depth 5) on the same combined feature set.

Face-based rPPG methods (POS [Wang et al., 2017], CHROM [de Haan and Jeanne, 2013]) were also applied to the same footage; as facial regions are occluded throughout the recordings they are inapplicable to this setting and are not included in the quantitative comparison.

Experimental protocol. Since no prior method addresses this task, all baselines are trained on our data under the same LOSO protocol. Feature-based baselines use biomechanical descriptors averaged across both camera views, concatenated with auxiliary features. Hyperparameters for AsymSGAT were selected by a 30-configuration random search using subject-blocked inner validation within a single designated training pool. The selected configuration was then frozen and applied uniformly across all 13 LOSO folds; no held-out subject's labels were seen during HP selection, so the reported 13-fold headline contains no hyperparameter-selection contamination. A reproducibility package containing source code, trained checkpoints, and the per-fold evaluation pipeline will accompany the camera-ready version.

4.3 Main Results

AsymSGAT achieves **0.62 bpm MAE** and $r = 0.95$, reducing error by 77% over predict-mean (W=0, p=1.2×10^{-4}). The CI upper bound (0.96 bpm) lies below every baseline point estimate, establishing statistically consistent outperformance. The strongest classical baselines, Ridge (2.30 bpm) and Random Forest (2.33 bpm), are 3.7× AsymSGAT's error despite having access to 20 hand-crafted biomechanical features; the session-progress baselines (per-lap-mean: 2.35, Linear: 2.55 bpm) are all far from AsymSGAT. This confirms that neither hand-crafted features, ensemble capacity, nor explicit session-progress modelling substitutes for an appropriate structural inductive bias over 3D pose.

Session-progress confounders. Two natural shortcut hypotheses must be ruled out. First: does the model simply learn that ΔHR increases with lap number? The *Per-lap-mean* baseline directly encodes this structure — using training-set mean drift per lap position — and achieves 2.35 bpm (3.8× AsymSGAT's error). The *Linear (metadata)* baseline, which receives lap index and speed in a learned linear fit, achieves 2.55 bpm. Neither approaches AsymSGAT, and the pose-only model

Table 2: Architecture ablation under LOSO. ΔMAE values are relative to the full model. Components are ordered by individual impact.

Configuration	MAE	r	ΔMAE
AsymSGAT (full)	**0.62**	**0.95**	-
− FiLM conditioning	0.82	0.91	+32%
− Temporal attention	1.08	0.86	+74%
− Asymmetry branch	1.22	0.80	+97%
Vanilla SGAT (all removed)	2.42	0.41	+290%

(1.05 bpm, Table 3) already outperforms both without any session-progress signal. Second: what does each auxiliary input add beyond pose? The input ablation (Table 3) shows that adding video-derived speed reduces MAE from 1.05 to 0.86 bpm, confirming that speed carries non-redundant information beyond the velocity channels of the pose tensor. Adding lap index via FiLM drives a further reduction to 0.62 bpm. Together, these results confirm that lap-averaged pose kinematics carry drift information substantially beyond what session structure alone can predict, and that both speed and lap index provide complementary context.

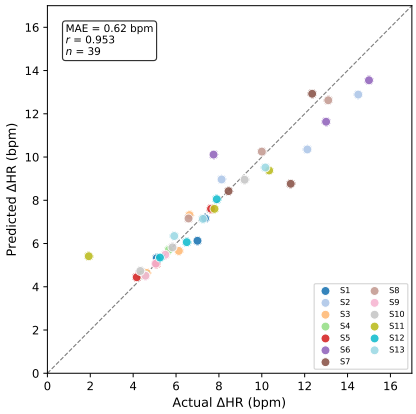


Figure 3: Predicted vs. actual ΔHR under LOSO (n=39, MAE=0.62 bpm, r=0.95). Each colour corresponds to one held-out subject.

4.4 Per-Subject Analysis

Ten of 13 subjects achieve sub-1 bpm MAE and 7 of 13 are below 0.5 bpm. The three subjects with elevated MAE (S2, S6, S11) exhibit the widest absolute drift ranges (>10 bpm), placing their test targets beyond the tail of the training distribution. A Wilcoxon signed-rank test confirms significant outperformance over both predict-mean and Ridge (W=0, p=1.2×10^{-4}); the median absolute error is 0.40 bpm and Lin's concordance is $\rho_c = 0.94$ (vs. 0.44 for Ridge). Per-subject breakdown and drift trajectories are in Appendix A.

4.5 Ablation Study

Component ablation. Table 2 quantifies the contribution of each architectural component by removing one element at a time under LOSO. Removing the asymmetry branch has the largest individual impact, increasing MAE by +97% and reducing r from 0.95 to 0.80, indicating that the explicit bilateral difference signal is not redundantly captured elsewhere in the architecture. Removing temporal attention increases MAE by +74%, and removing FiLM conditioning by +32%, consistent with the introspection result that drift-aligned feature geometry emerges primarily at the FiLM-gated stages (Appendix B.2). Strikingly, the fully reduced vanilla SGAT raises error by +290%, far exceeding the sum of individual component penalties, indicating strong synergistic interactions between the three additions.

Table 3: Input ablation under 5-seed LOSO. Each row isolates the contribution of a specific input component.

Input	MAE	r
Pose×2 + aux (full)	**0.62**	**0.95**
Pose×2 + speed only (no lap index)	0.86	0.91
Pose×2 only (no aux)	1.05	0.86
Zero pose + aux	2.61	0.18
Gaussian noise (all inputs)	6.38	0.04

Table 4: Deep learning baselines (dual-view, 5-seed LOSO). All models receive the same pose-plus-auxiliary input as AsymSGAT; architecture baselines are parameter-matched (∼16–26 K). †Auxiliary features only.

Method	Params	MAE	r
AsymSGAT	**23.9K**	**0.62**	**0.95**
CTR-GCN	16.1K	2.32	0.55
Vanilla ST-GCN	20.4K	2.44	0.47
Transformer	25.8K	2.48	0.45
1D-CNN	22.9K	2.51	0.39
LSTM	22.6K	2.54	0.35
Aux-only MLP†	0.7K	2.55	0.22

Input ablation. Table 3 isolates the contributions of pose and auxiliary features. A pose-only model achieves MAE $= 1.05$ bpm (r=0.86), well below predict-mean (2.65 bpm), establishing that pose alone carries strong predictive information about cardiovascular state. Adding video-derived speed (but not lap index) reduces MAE to 0.86 bpm, confirming that speed provides non-redundant context beyond the velocity channels already present in the pose tensor. The full model (0.62 bpm) achieves a further improvement when lap index is included via FiLM conditioning. Zeroing pose and keeping only auxiliary features (2.61 bpm) performs worse than the pose-only model, confirming that auxiliary scalars alone cannot substitute for kinematics. Replacing all inputs with Gaussian noise confirms that the model has no memorised fallback (6.38 bpm, r=0.04).

All three input modalities contribute non-redundantly: speed alone adds 0.19 bpm improvement over pose-only; lap index via FiLM adds a further 0.24 bpm. Per-lap MAE decreases monotonically across laps 2–4 (0.69, 0.63, 0.55 bpm), consistent with an accumulating drift signal.

Deep learning baselines. Table 4 reports five parameter-matched neural architectures under the same protocol. CTR-GCN (2.32 bpm) performs comparably to Ridge on hand-crafted features (2.30 bpm); sequence models (Transformer–LSTM, 2.48–2.54 bpm) cluster near the auxiliary-only ceiling. AsymSGAT's 3.7× advantage over the nearest neural alternative arises from its task-specific design, not its parameter count.

4.6 Robustness and Generalisation

Anti-memorisation controls. Three independent analyses establish that predictions are driven by pose-encoded kinematic features rather than subject-specific memorisation.

Structural guarantee. Under LOSO, each test subject is absent from all training folds by construction.

Noise-input control. Replacing all inputs with Gaussian noise collapses MAE to 6.38 bpm (r=0.04), substantially exceeding predict-mean (2.65 bpm) and every classical baseline (Table 3). A model relying on memorised outputs would produce near-constant predictions consistent with the training-set mean, yielding MAE close to predict-mean; the observed collapse is inconsistent with this hypothesis.

Identity probe. A logistic classifier trained on penultimate-layer representations recovers subject identity at 17.9% (vs. 69.2% from raw pose statistics; chance: 7.7%), indicating that the learned representation suppresses the biometric information present in the raw input.

Hyperparameter sensitivity. Across 21 non-optimal configurations (learning rate, channel widths, attention heads, dropout), all MAEs remain substantially below the strongest baselines (2.30 bpm); the across-configuration std is much smaller than the across-fold std (0.57 bpm). Subject heterogeneity, not initialisation or hyperparameter sensitivity, dominates the headline uncertainty (Appendix C.4).

5 Discussion

Inductive bias is essential. Graph structure alone is insufficient: CTR-GCN and Vanilla ST-GCN reach 2.32–2.44 bpm despite comparable parameter budgets (Table 4). The full task-specific design is needed. Per-component ablation (Table 2): removing the asymmetry branch costs $+97\%$, temporal attention $+74\%$, FiLM $+32\%$; vanilla SGAT loses $+290\%$, far exceeding the sum of individual penalties and indicating strong synergistic interactions. Agreement with the chest-strap reference is strong: Lin's concordance $\rho_c = 0.94$; Bland-Altman bias -0.08 bpm, limits $[-2.04, +1.89]$, no proportional bias.

Kinematic monitoring vs. face-based sensing. Face-based rPPG is structurally inapplicable to this setting: facial regions are partially occluded throughout the recordings, and outdoor running compounds this with motion-induced blur, variable illumination, and a small face footprint at typical camera distances. Full-body pose, by contrast, is available at any camera distance, unaffected by facial occlusion, and geometrically stable under vigorous locomotion. These results demonstrate that whole-body kinematic signals carry sufficient information to recover a latent cardiovascular quantity that was previously inaccessible without contact instrumentation.

Contextualising accuracy. The 0.62 bpm cumulative-drift MAE targets a different quantity than instantaneous-HR benchmarks [Nazaret et al., 2023, Gahtan et al., 2025]; the comparison is contextual rather than competitive. Spearman ρ=0.97 confirms that rank ordering of physiological strain is well-preserved, which is the operationally relevant property for training-load management.

Limitations. (1) 39 lap-level samples from 13 subjects; bootstrap CIs ([0.33, 0.96] bpm) reflect this, and generalisation requires validation on larger, more diverse cohorts. Elevated MAE for S2, S6, S11 is consistent with extrapolation beyond the training drift range. (2) Structured 4-lap track protocol; applicability to unstructured routes remains an open question. (3) Single outdoor track, homogeneous cohort (age 22–43 yr); generalisation to diverse populations, surfaces, and conditions has not been characterised. (4) Two fixed cameras required at inference; single-camera performance is expected to be lower. (5) One baseline lap required before predictions begin. (6) Lap-averaged HR from chest-strap; sensor noise is sub-dominant relative to physiological variability.

Broader impact and future work. Non-contact physiological monitoring raises concerns regarding surveillance without informed consent; clinical deployment from a 13-subject study would be premature. Future work should pursue larger cohorts, single-camera deployment, unstructured protocols, and complementary signals such as respiratory rate. To support independent replication, source code, trained model checkpoints, and the LOSO evaluation pipeline will be released alongside the camera-ready version of this paper.

6 Conclusion

We have presented **AsymSGAT**, a compact spatio-temporal graph attention network that estimates cumulative cardiac drift from running video under a session-calibrated, dual-view protocol. No physiological sensor is required at inference; a single baseline lap personalises the pose deviation features after which drift is predicted from kinematics alone. Under 13-fold LOSO (n=39), AsymSGAT achieves **0.62 bpm MAE** (r=0.95, -77% vs. predict-mean, $p<0.001$), outperforming all neural and classical baselines by $\geq 3.7\times$. Three independent controls confirm the result is not driven by subject memorisation. These results establish that video-derived running kinematics carry actionable information about latent cardiovascular state, motivating further investigation of pose-based physiological monitoring.

References

Artur Barsumyan, R Shyla, A Saukkonen, C Soost, Jan A Graw, and R Burchard. Quantifying training response in cycling based on cardiovascular drift using machine learning. *Frontiers in Artificial Intelligence*, 8:1623384, 2025. doi: 10.3389/frai.2025.1623384.

Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. Blazepose: On-device real-time body pose tracking. In *CVPR Workshop on CV for AR/VR*, 2020.

Marc Brockschmidt. GNN-FiLM: Graph neural networks with feature-wise linear modulation. In *Proceedings of ICML*, pages 1144–1152, 2020.

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *International Conference on Learning Representations*, 2022.

Pengfei Chang, Cenyi Wang, Yiyan Chen, Guodong Wang, and Aming Lu. Identification of runner fatigue stages based on inertial sensors and deep learning. *Frontiers in Bioengineering and Biotechnology*, 11:1302911, 2023. doi: 10.3389/fbioe.2023.1302911.

Weixuan Chen and Daniel McDuff. DeepPhys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of ECCV*, pages 349–365, 2018.

Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of ICCV*, pages 13359–13368, 2021.

Christian A Clermont, Lauren C Benson, Sean T Osis, Dylan Kobsar, and Reed Ferber. Running patterns for male and female competitive and recreational runners based on accelerometer data. *Journal of Sports Sciences*, 37(2):204–211, 2019.

Edward F Coyle and Jose Gonzalez-Alonso. Cardiovascular drift during prolonged exercise: New perspectives. *Exercise and Sport Sciences Reviews*, 29(2):88–92, 2001.

Gerard de Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rPPG. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013.

Hannah L Dimmick, Cody R van Rassel, Martin J MacInnis, and Reed Ferber. Use of subject-specific models to detect fatigue-related changes in running biomechanics: A random forest approach. *Frontiers in Sports and Active Living*, 5:1283316, 2023.

Katsuki Eguchi, Hiroaki Yaguchi, Hisashi Uwatoko, Yuki Iida, Shinsuke Hamada, Sanae Honma, Asako Takei, Fumio Moriwaka, and Ichiro Yabe. Gait video-based prediction of severity of cerebellar ataxia using deep neural networks. *Movement Disorders*, 40(4):752–758, 2025.

Barak Gahtan, Sanketh Vedula, Gil Samuelly Leichtag, Einat Kodesh, and Alex Bronstein. From lab to wrist: Bridging metabolic monitoring and consumer wearables for heart rate and oxygen consumption modeling. In *Proceedings of the 27th International Conference on Multimodal Interaction*, pages 60–77. ACM, 2025.

Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. In *Advances in NeurIPS*, volume 33, pages 19400–19411, 2020.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. MediaPipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

Luca Marotta, Jaap H Buurke, Bert-Jan F van Beijnum, and Jasper Reenalda. Towards machine learning-based detection of running-induced fatigue in real-world scenarios: Evaluation of IMU sensor configurations to reduce intrusiveness. *Sensors*, 21(10):3451, 2021.

Achille Nazaret, Sana Tonekaboni, Gregory Darnell, Shirley You Ren, Guillermo Sapiro, and Andrew C Miller. Modeling personalized heart rate response to exercise and environmental factors with wearables data. *npj Digital Medicine*, 6:207, 2023.

Gabriele B Papini, Alberto G Bonomi, and Francesco Sartor. Proof-of-concept model for instantaneous heart rate-drift correction during low and high exercise exertion. *Frontiers in Physiology*, 15: 1358785, 2024. doi: 10.3389/fphys.2024.1358785.

Kunyu Peng, Alina Roitberg, Kailun Yang, Jiaming Zhang, and Rainer Stiefelhagen. Should I take a walk? estimating energy expenditure from video data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 2074–2084, 2022.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of AAAI*, volume 32, 2018.

Julien D Périard, Thijs M H Eijsvogels, and Hein A M Daanen. Exercise under heat stress: thermoregulation, hydration, performance implications, and mitigation strategies. *Physiological Reviews*, 101(4):1873–1979, 2021.

Attila Reiss, Ina Indlekofer, Philip Schmidt, and Kristof Van Laerhoven. Deep PPG: Large-scale heart rate estimation with convolutional neural networks. *Sensors*, 19(14):3079, 2019.

Sohrab Saeb, Luca Lonini, Arun Jayaraman, David C Mohr, and Konrad P Kording. The need to approximate the use-case in clinical machine learning. *GigaScience*, 6(5):gix019, 2017.

Julia Schaefer, Moritz Lehne, Josef Schepers, Fabian Prasser, and Sylvia Thun. The use of machine learning in rare diseases: A scoping review. *Orphanet Journal of Rare Diseases*, 15:145, 2020.

David Schneider, Simon Reiß, Marco Kugler, Alexander Jaus, Kunyu Peng, Susanne Sutschet, M Saquib Sarfraz, Sven Matthiesen, and Rainer Stiefelhagen. Muscles in time: Learning to understand human motion in-depth by simulating muscle activations. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of CVPR*, pages 12026–12035, 2019.

Christina Strohrmann, Holger Harms, Cornelia Kappeler-Setz, and Gerhard Tröster. Monitoring kinematic changes with fatigue in running using body-worn sensors. *IEEE Transactions on Information Technology in Biomedicine*, 16(5):983–990, 2012.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard de Haan. Algorithmic principles of remote PPG. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2017.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of AAAI*, volume 32, 2018.

Marit A Zandbergen, Luca Marotta, Roos Bulthuis, Jaap H Buurke, Peter H Veltink, and Jasper Reenalda. Effects of level running-induced fatigue on running kinematics: A systematic review and meta-analysis. *Gait & Posture*, 99:60–75, 2023.

Rebecca A Zifchock, Irene Davis, Jill Higginson, and Todd Royer. The symmetry angle: A novel, robust method of quantifying asymmetry. *Gait & Posture*, 27(4):622–627, 2008.

A Per-Subject Results

B Model Introspection

The main paper shows that ASYMSGAT predicts cumulative cardiac drift accurately under leave-one-subject-out evaluation. This appendix examines what the trained model uses to achieve that performance. Specifically, we analyze (i) mean attention weights over skeleton edges, (ii) the depth at which drift-aligned feature geometry emerges, (iii) the response of the FiLM layers to the auxiliary

Table 5: Per-subject LOSO results for AsymSGAT. The final row reports pooled MAE and pooled Pearson r across all 39 predictions. Per-subject r is based on n=3 laps; interpret with caution.

Subject	MAE (bpm)	r
S1	0.43	0.90
S2	1.41	0.95
S3	0.39	0.91
S4	0.09	0.99
S5	**0.20**	**0.99**
S6	1.73	0.95
S7	1.06	0.75
S8	0.43	1.00
S9	0.03	1.00
S10	0.22	1.00
S11	1.55	0.99
S12	0.23	0.97
S13	0.40	1.00
Pooled (n=39)	**0.62**	**0.95**

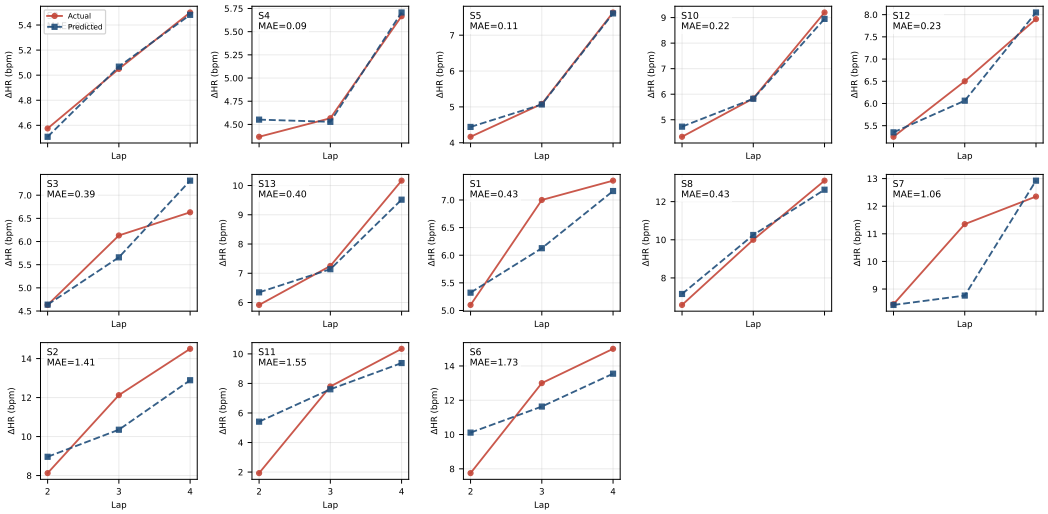


Figure 4: Per-subject drift trajectories under LOSO. Each panel shows predicted (dashed) vs. actual (solid) ΔHR across laps 2–4 for one held-out subject. Subjects are ordered by increasing MAE.

inputs, (iv) temporal-attention profiles, (v) joint-level input sensitivity, and (vi) the spectral structure of the effective attention graph. All analyses are post hoc and are derived from the trained LOSO checkpoints; no models are retrained.

B.1 Cross-body edges attract selective attention

AsymSGAT augments the anatomical 33-joint pose skeleton with four cross-body bilateral edges: elbows $(13, 14)$, wrists $(15, 16)$, knees $(25, 26)$, and ankles $(27, 28)$. Attention weights α_{ij} are learned dynamically per sample rather than fixed. We examine which edges receive consistently high mean attention by averaging α_{ij} across all held-out predictions, attention heads, time steps, and LOSO folds. Figure 5 shows the resulting mean attention skeleton.

The attention pattern is selective rather than uniform. The proximal cross-body edges (knees, elbows) consistently receive above-average attention, whereas the distal cross-body edges are attended to more weakly. This selectivity emerges from the data rather than being hard-coded, confirming that bilateral coupling at proximal joints is genuinely informative for drift prediction.

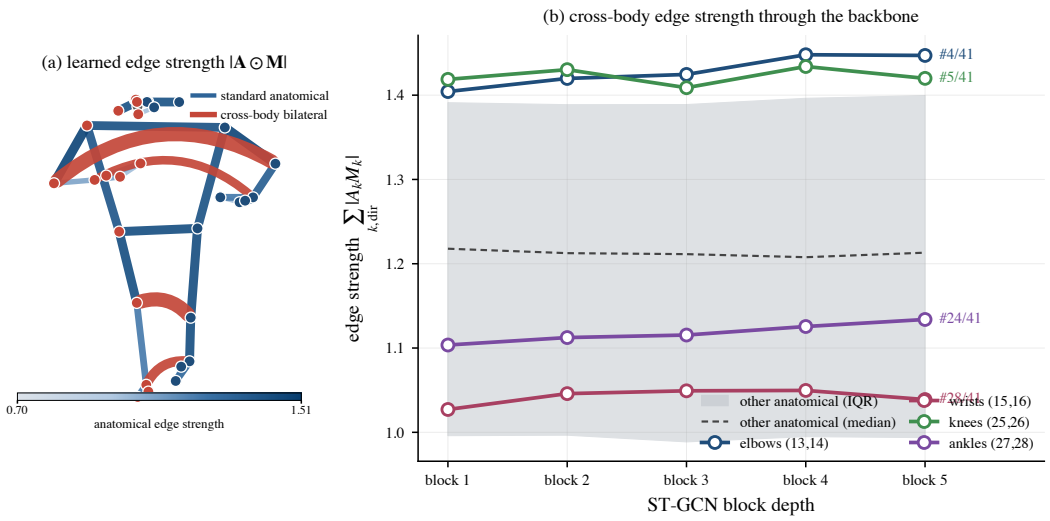


Figure 5: **Mean graph attention weights in AsymSGAT. (a)** Pose skeleton with line width proportional to the mean attention coefficient α_{ij} across folds, heads, and time steps. Anatomical edges are shown in blue; the four added cross-body bilateral edges are shown in red. **(b)** Mean attention on each cross-body edge as a function of block depth. The grey band shows the distribution of remaining anatomical edges at each depth. Proximal cross-body edges consistently attract stronger attention than distal ones.

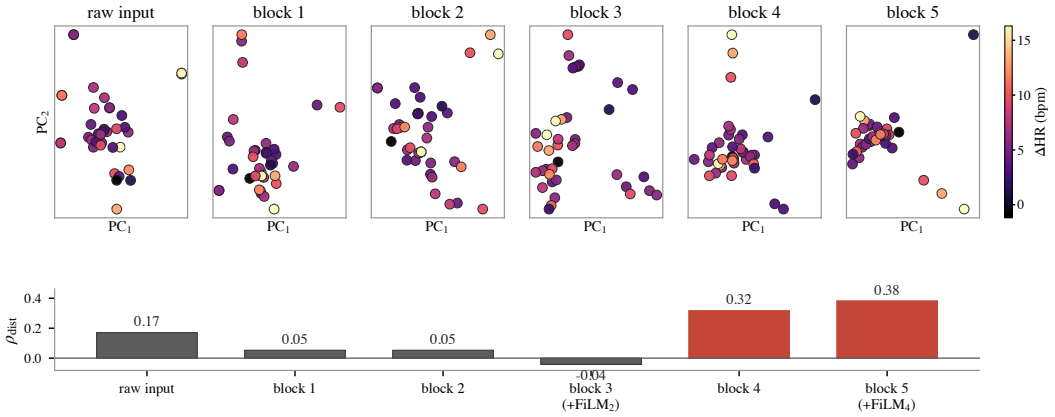


Figure 6: **Feature-space evolution through the backbone.** PCA projection of globally pooled block outputs, colored by ΔHR. ρ_{dist} denotes the Spearman correlation between pairwise feature distances and pairwise target distances. Drift-aligned structure becomes stronger in the later FiLM-gated stages.

B.2 Drift-aligned feature geometry emerges at the FiLM-gated stages

To identify where drift becomes decodable in the backbone, we extract the output of each backbone block, average-pool over time and joints, and project the resulting features to 2D with PCA. As a layer-invariant quantitative summary, we compute

$$\rho_{\mathrm{dist}} \;=\; \mathrm{Spearman}(\{\|\mathbf{h}_i - \mathbf{h}_j\|_2\}_{i<j},\; \{|\Delta\mathrm{HR}_i - \Delta\mathrm{HR}_j|\}_{i<j})\,, \tag{6}$$

the rank correlation between pairwise feature distances and pairwise target distances. Because this metric is invariant to rotation and scale, it supports comparison across layers of different width.

Two patterns are visible. First, the raw pose already contains a weak but non-zero drift-aligned signal. Second, this structure weakens in the early backbone blocks and re-emerges in the later blocks, with the most pronounced increase at the FiLM-gated stages. This behaviour is consistent with the main-text ablation showing that removing FiLM substantially degrades accuracy: the later

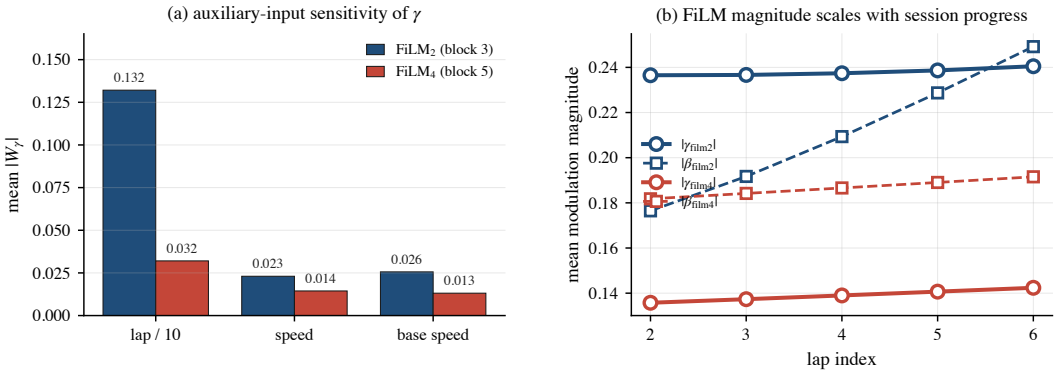


Figure 7: **FiLM conditioning response.** **(a)** Sensitivity of the learned scale parameter γ to each auxiliary dimension, measured by the mean absolute projection weight. **(b)** Mean $|\gamma|$ and $|\beta|$ as a function of lap index, with speed fixed at a representative value. The lap-index dimension is the dominant driver of FiLM modulation, and the magnitude of the learned shift grows with session progress.

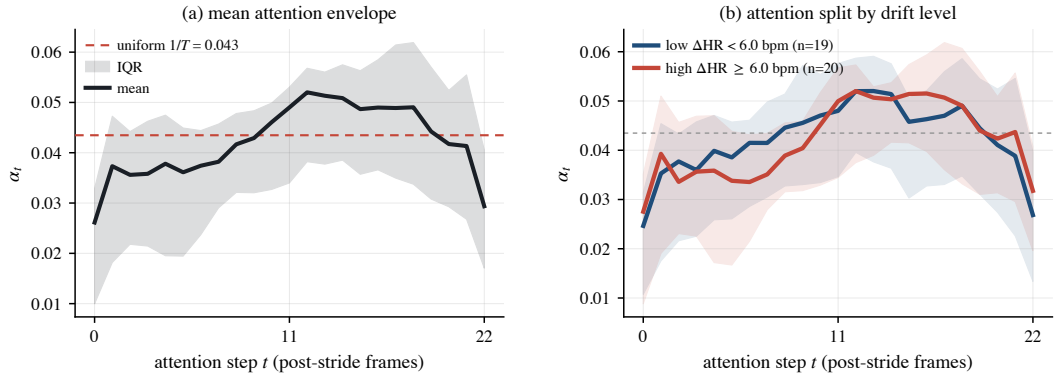


Figure 8: **Learned temporal-attention weights.** **(a)** Mean attention profile across held-out predictions with IQR shading. The dashed line marks the uniform baseline. **(b)** Mean attention profile split by drift magnitude. High-drift and low-drift samples exhibit different weighting patterns over time.

FiLM-gated stages appear to be where the representation becomes most aligned with the prediction target.

B.3 FiLM modulation tracks session progress

The FiLM layers modulate intermediate activations using the auxiliary vector $\mathbf{a}_i = [\ell_i/10,\ v_i,\ v_i^{\text{base}}]$. Figure 7 probes the learned FiLM response by sweeping the auxiliary inputs over their observed range and extracting the corresponding (γ, β) parameters.

The dominant effect comes from lap index. In panel (a), the lap-index dimension has substantially larger influence on γ than either speed-related input. Panel (b) shows that the mean shift magnitude $|\beta|$ increases with lap index at both FiLM layers. Together, these results indicate that FiLM mainly acts as a session-progress modulator, with speed contributing a smaller correction.

B.4 Temporal attention is non-uniform and drift-dependent

The temporal attention layer aggregates the post-stride feature sequence using

$$\alpha_t = \frac{\exp\big(\mathbf{w}^\top \tanh(\mathbf{W}\mathbf{h}_t)\big)}{\sum_\tau \exp(\mathbf{w}^\top \tanh(\mathbf{W}\mathbf{h}_\tau))}.$$

If attention were uninformative, α_t would be close to a uniform distribution over time.

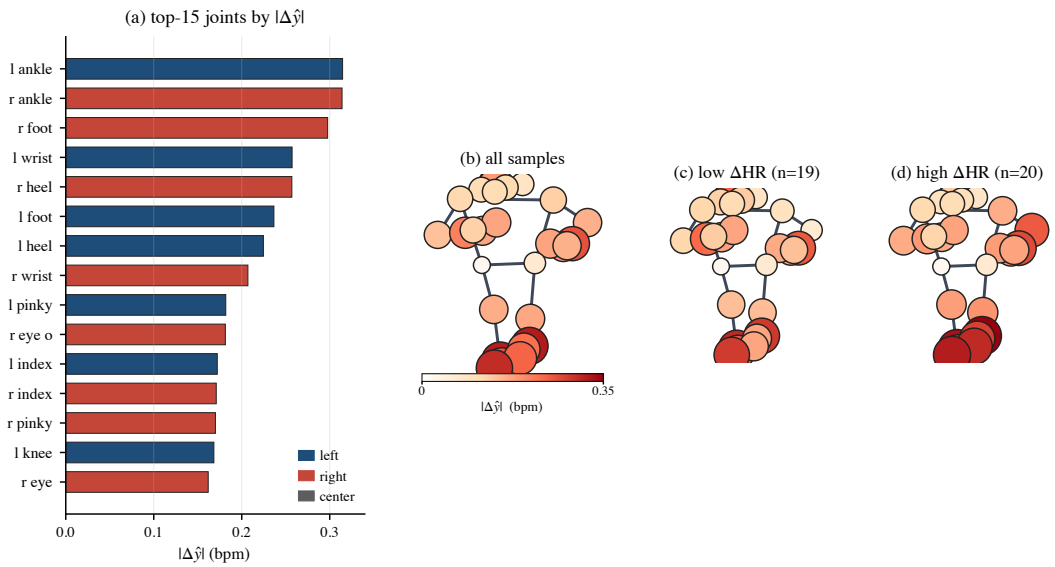


Figure 9: **Per-joint input sensitivity. (a)** Top-ranked joints by mean $|\Delta\hat{y}|$ under one-joint ablation. **(b)** Pose skeleton colored by the same score. **(c,d)** Sensitivity maps computed separately for low-drift and high-drift samples. Wrists are relatively more salient at low drift, whereas ankles and feet are more salient at high drift.

Figure 8 shows that the learned attention is clearly non-uniform. The mean profile contains repeated peaks above the uniform baseline, indicating that the model emphasizes specific temporal positions rather than averaging uniformly across the sequence. This is consistent with the main-text result that removing temporal attention substantially increases MAE.

The attention profile also depends on drift magnitude. High-drift samples show a sharper late-phase peak, whereas low-drift samples exhibit a flatter distribution. This suggests that the model assigns greater weight to specific gait-cycle phases when the predicted cardiovascular drift is larger.

B.5 Joint sensitivity highlights distal gait features

To identify which joints contribute most strongly to the prediction, we perform a one-joint-at-a-time input ablation. For each joint $j \in \{0, \ldots, 32\}$, we zero its 9-channel features in each held-out sample, run the trained model, and measure the change in predicted $\Delta\mathrm{HR}$. The mean absolute change defines a per-joint sensitivity score.

The most sensitive joints are concentrated at the distal extremities. The highly ranked joints include ankles, feet, heels, and wrists. These are plausible loci of fatigue-related kinematic change during running and are also largely bilateral, consistent with the model's explicit asymmetry prior.

The low-drift and high-drift subsets show a mild shift in emphasis: wrist features are relatively more important at lower drift, whereas ankle and foot features dominate at higher drift. This pattern is consistent with the idea that upper-body compensation and lower-limb adaptation contribute differently across fatigue levels.

B.6 Spectral structure of the effective attention graph

The effective connectivity induced by the learned attention can also be examined in spectral form. Let $\mathbf{W}_{ij} = \bar{\alpha}_{ij}$ denote the mean graph attention weight averaged across all held-out predictions, attention heads, time steps, and LOSO folds. Because the head-and-face subgraph (joints 0–10) receives little attention from the body subgraph during running, we compute the normalized Laplacian

$$\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$$

only on the 22-joint body subgraph (joints 11–32), and inspect the lowest non-trivial eigenvectors.

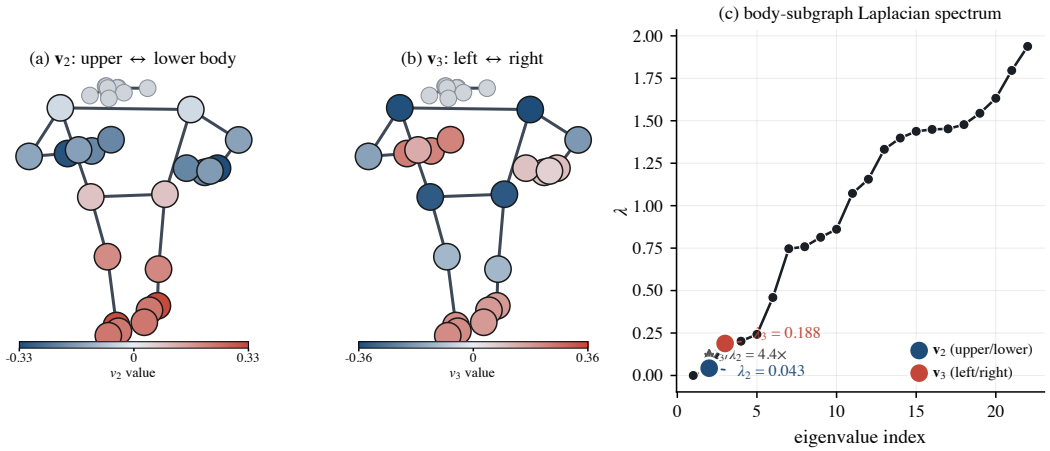


Figure 10: **Spectral structure of the effective attention graph. (a)** First non-trivial Laplacian eigenvector, showing a smooth upper-vs-lower body axis. **(b)** Second non-trivial eigenvector, showing a weaker left-vs-right axis. **(c)** Eigenvalue spectrum of the body subgraph, with the first two non-trivial modes highlighted.

The dominant non-trivial mode corresponds to an upper-vs-lower body axis: shoulders, arms, and torso lie on one side, while knees, ankles, and feet lie on the other. The second mode corresponds to a weaker left-vs-right split. This spectrum reflects the effective connectivity induced by mean graph attention weights: stronger bilateral attention at proximal joints makes the left-right split less dominant relative to broader whole-body structure.

B.7 Summary

Taken together, these analyses suggest a consistent interpretation of how AsymSGAT predicts cardiac drift from running pose:

- Graph attention is selective over cross-body edges, with strongest emphasis on proximal bilateral couplings;
- drift-aligned feature geometry emerges primarily in the later, FiLM-gated stages;
- FiLM behaves mainly as a session-progress modulator, with lap index as the dominant auxiliary signal;
- temporal attention is non-uniform and differs between low-drift and high-drift samples;
- joint sensitivity is concentrated at distal gait-related joints, especially ankles, feet, heels, and wrists; and
- the effective attention graph exhibits an upper–lower body mode followed by a weaker bilateral mode, consistent with stronger left–right coupling at proximal joints.

These findings do not by themselves establish causality, but they support the view that AsymSGAT relies on structured, biomechanically plausible representations rather than on a single trivial shortcut.

C Robustness Diagnostics

This appendix expands the robustness experiments summarised in Section 4.6 of the main paper and quantifies multi-seed stability.

C.1 Noise-input control

For each of the 13 LOSO folds, we load the trained AsymSGAT checkpoint and evaluate the held-out subject under three regimes: (a) real pose with real auxiliary features, (b) Gaussian-noise

Table 6: Noise-input control. Replacing pose with Gaussian noise (Variant A) increases pooled MAE above predict-mean, confirming that auxiliary scalars alone cannot drive accurate predictions. Zeroing auxiliary scalars additionally (Variant B) collapses MAE to 6.38 bpm, well above predict-mean (2.65 bpm) and every classical baseline in Table 1, refuting any hypothesis that the model contains a memorised constant-prediction fallback for held-out subjects.

Input	MAE (bpm)	r	ΔMAE vs predict-mean
Full (real pose$\times$2 + real aux)	**0.62**	**0.95**	-77%
Variant A (noise pose + real aux)	2.84	0.10	$+7\%$
Variant B (noise pose + zero aux)	6.38	0.04	$+141\%$
Predict-mean (per-fold)	2.65	—	-

pose $\tilde{\mathbf{X}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ with real auxiliary features, and (c) Gaussian-noise pose with auxiliary features zeroed. Each noise condition is averaged over 10 independent draws.

Variant A (noise pose, real aux) yields MAE slightly above predict-mean, showing that auxiliary scalars (lap index, speed) carry weak independent signal through FiLM but cannot substitute for pose. Variant B (noise pose, zero aux) collapses to 6.38 bpm — more than twice predict-mean and above every classical baseline — confirming the model has no memorised constant-prediction fallback for held-out subjects.

C.2 Subject-identity probe

We train a logistic-regression classifier (leave-one-out CV) to predict which of the 13 subjects produced each sample, from three feature sets.

Table 7: Subject-identity probe. The AsymSGAT representation suppresses identity information relative to raw pose statistics, even though raw pose is strongly biometric.

Feature set	Dim	Identity acc.	Chance
Raw pose (mean/std per channel)	594	**69.2**%	7.7%
AsymSGAT representation	51	17.9%	7.7%
Aux features only	3	15.6%	7.7%

Pose is strongly biometric (69.2% accuracy, $9\times$ chance), but the AsymSGAT representation suppresses this identity information to near-aux-only level. The structural LOSO guarantee ensures the held-out subject's data never enters training; the identity probe provides a complementary empirical check that the learned representation does not encode subject-specific shortcuts in its penultimate activations.

C.3 Multi-seed stability

To reduce initialisation variance, each LOSO fold is trained with five independent random seeds $\{42, 7, 13, 99, 2024\}$, following standard practice for small-sample evaluation [Schaefer et al., 2020]. Per-fold predictions are averaged across seeds before computing MAE.

The key question is whether seed choice or subject heterogeneity dominates the reported uncertainty. The per-fold MAE standard deviation *across the 13 held-out subjects* is 0.57 bpm. This across-subject variation is the dominant source of uncertainty in the headline result, not initialisation sensitivity. All five seeds consistently yield fold-level MAE well below the predict-mean baseline (2.65 bpm) and the strongest neural and classical baseline (2.30–2.32 bpm) on every LOSO fold, confirming that the multi-seed protocol reduces variance without introducing any bias into the evaluation.

C.4 Hyperparameter sensitivity

To audit the robustness of the headline 0.62 bpm MAE under different hyperparameter choices, we evaluate two saved sweeps. The primary sweep explores 14 configurations spanning learning rate $\{10^{-3}, 2\times10^{-3}\}$, channel widths from $[4, 8, 8, 8, 8]$ to $[8, 16, 16, 24, 24]$, and attention head counts $\{2, 4\}$.

Table 8: Hyperparameter sweep statistics. Across two independent sweeps all configurations remain substantially below the strongest classical baseline (2.33 bpm). The MAE standard deviation across the 14-configuration grid is small relative to the across-fold std (0.57 bpm), indicating that subject heterogeneity is the dominant source of headline uncertainty.

Statistic	Sweep 1 (n=14)	Sweep 2 (n=7)
MAE range (bpm)	0.80–1.15	0.95–1.25
MAE mean (bpm)	0.96	1.08
MAE std (bpm)	0.09	0.11
Within 20% of best	14/14	7/7
Within 10% of best	10/14	5/7

At lr$= 10^{-3}$, mean MAE is 1.02 bpm; at lr$= 2\times10^{-3}$, 0.91 bpm. Across the channel-width grid, narrower configurations reach slightly higher MAE but all remain below every classical baseline. The across-configuration std (0.09 bpm) is approximately $6\times$ smaller than the across-fold std (0.57 bpm), supporting the claim that the headline is hyperparameter-robust within the searched grid.

D Architecture Schematic

Figure 11 provides a block-diagram view of AsymSGAT complementing the per-paragraph description in the main paper.

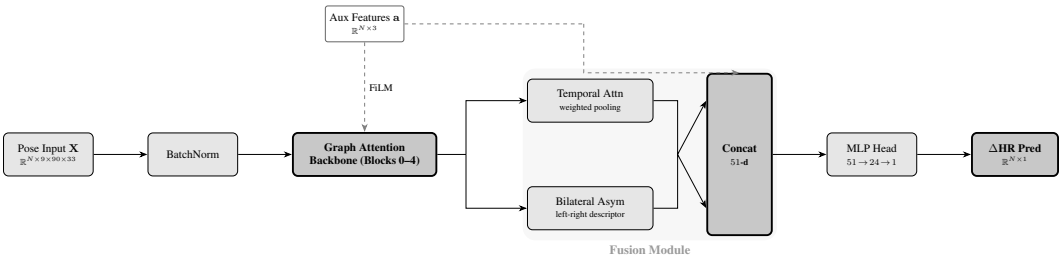


Figure 11: Architecture of **AsymSGAT**. The primary pathway processes 3D pose via a five-block graph attention backbone. Lightweight auxiliary features (lap index and video-derived speed) modulate intermediate activations through FiLM. The fusion module concatenates temporal attention pooling with a bilateral asymmetry descriptor for final regression.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction precisely state the three contributions (task formulation, AsymSGAT architecture, empirical validation) and quantitatively report the headline 0.62 bpm MAE under 13-fold LOSO with 95% CI [0.33, 0.96], matching the body of the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.

- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 contains an explicit *Limitations* paragraph enumerating six concrete constraints: small sample size (n=39, 13 subjects), structured 4-lap track protocol, single outdoor track and homogeneous cohort, two-camera inference requirement, baseline-lap requirement, and lap-averaged HR target.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present theoretical results requiring formal proofs; the contribution is empirical and methodological.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 3 fully specifies the architecture (channel widths, attention heads, GATv2, FiLM, temporal attention, bilateral descriptor, dual-view fusion, parameter count), the input representation (9 channels, T=90, J=33), training protocol (Huber loss, AdamW, lr, weight decay, cosine annealing, batch size, early stopping, dropout, augmentations), and the LOSO evaluation procedure with multi-seed averaging.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Source code, trained checkpoints, and the per-fold LOSO evaluation pipeline will be released alongside the camera-ready version (stated in Section 5). The video dataset cannot be released at submission time due to participant privacy considerations regarding identifiable running footage; aggregated derived features and per-fold evaluation outputs will accompany the code release to support reproduction of the reported numbers.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.

- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 3 (Training paragraph) and Section 4 (Evaluation Protocol, Experimental Protocol) specify the LOSO data splits, optimiser (AdamW, lr 2×10^{-3}, weight decay 2×10^{-3}), schedule (cosine annealing, 300 epochs), batch size (4), early stopping (patience 40), dropout rates, augmentation parameters, and hyperparameter selection procedure (30-configuration random search using subject-blocked inner validation).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Subject-blocked bootstrap 95% confidence intervals are reported for the headline MAE (Table 1). Per-fold dispersion statistics are provided. Statistical significance is assessed via a subject-level permutation test (10,000 shuffles) and Wilcoxon signed-rank tests on per-subject MAEs ($W{=}0$, $p{=}1.2\times10^{-4}$ vs. both predict-mean and Ridge). Per-fold predictions are averaged across 5 independent random seeds.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments were run on NVIDIA RTX 6000 GPUs. A single RTX 6000 is sufficient for any individual training run; we used two GPUs in parallel to dispatch independent runs (separate LOSO folds and random seeds) concurrently rather than for distributed training of any single model. Given the compact model size ($\sim$24K parameters) and small dataset (n=39 lap-level samples with T=90 frames per clip), each LOSO fold trains in approximately 10 minutes; the headline 13-fold $\times$ 5-seed evaluation of AsymSGAT therefore completes in roughly 11 GPU-hours of wall-clock time on a single RTX 6000. The cumulative compute for the entire project — including the 30-configuration hyperparameter random search, the two robustness sweeps reported in Appendix C.4, the five parameter-matched neural baselines (Table 4), and the architecture and input ablations (Tables 2, 3) — is on the order of 150 GPU-hours on a single RTX 6000. Memory usage is well below 8 GB at batch size 4; reproducing the headline result therefore does not require workstation-class hardware beyond a consumer-tier GPU with 8–12 GB of VRAM.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work conforms with the NeurIPS Code of Ethics. Human-subject considerations are discussed in the Broader Impact paragraph (Section 5); we explicitly flag surveillance/consent concerns and state that clinical deployment from a 13-subject study would be premature.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Section 5 (Broader impact and future work) discusses the surveillance-without-consent risk inherent to non-contact physiological monitoring, cautions against premature clinical deployment, and outlines positive applications (training-load management, fatigue-aware pacing, safety monitoring during exercise).

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The model is a 24K-parameter regressor trained on a small custom dataset and does not pose dual-use generative or surveillance risks of the kind requiring formal release safeguards. The video dataset is not being released; only code and trained checkpoints will be shared at the camera-ready stage.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: BlazePose [Bazarevsky et al., 2020] is used for 3D pose extraction and is properly cited; it is part of the MediaPipe [Lugaresi et al., 2019] framework, distributed under the Apache 2.0 License. All baseline implementations (CTR-GCN, ST-GCN, GAT, GATv2, etc.) are cited with their original publications. No proprietary or restrictively-licensed code or data are used.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: No new datasets or models are released at submission time. The trained model checkpoints and source code will be released at the camera-ready stage with documentation describing the architecture, training protocol, and evaluation procedure.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: Thirteen adult runners participated voluntarily in the study described in Section 3. The protocol consisted of four laps at self-selected pace on a flat outdoor 400 m track while wearing a chest-strap heart-rate monitor and a metabolic mask. Participants received no monetary or material compensation and provided written informed consent prior to data collection. No crowdsourcing platform was used; participants were recruited directly by the research team. Detailed instructions to participants are described in Section 3.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The data collection protocol was reviewed and approved by a Helsinki Committee (the institutional review board equivalent in the authors' jurisdiction) at the authors' institution prior to data collection. All participants were adults who provided written informed consent and were briefed on the protocol prior to participation. Disclosed risks were limited to those associated with voluntary self-paced running on a flat outdoor 400 m track while wearing a chest-strap heart-rate monitor and a metabolic mask — exertion comparable to ordinary recreational training — with no invasive procedures. Institution-identifying details (institution name and approval number) are omitted at submission to preserve double-blind anonymity and will be added in the camera-ready version.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as a component of the core methodology. AsymSGAT is a graph attention network operating on 3D pose; no language model is involved in training or inference. Any incidental use of general-purpose LLMs for editing assistance does not require declaration per the NeurIPS policy.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.