Optical vs. Inertial Skeleton Tracking: A Positional Agreement Study for Sports Biomechanics

FitWise AI

September 24, 2026

Abstract

Reliable 3D skeleton tracking plays an important role in professional sports, allowing data analysts and performance coaches to detect deviations in movement biomechanics and prevent related negative impacts. This report assesses the agreement between our **optical** (single-camera, video-based) skeleton tracking and a **sensor-based** reference (an Xsens inertial motion-capture suit) on nine athletes performing three sports-relevant activities: steady treadmill running, shuttle runs, and a partial-occlusion drill in which a second person contests the tracked athlete. We stress that both systems are inherently error-prone. In optical systems, 2D-to-3D reconstruction is ambiguous by definition; in inertial systems, necessary biomechanical post-processing can introduce bias. We therefore report the **positional discrepancy** between the two systems, since neither can be treated as a gold standard for true joint-position estimation. After bringing the two systems into a common size and pose (the standard **PA-MPJPE** alignment, Appendix B), the median per-joint discrepancy is **2–5 cm** for the reliably comparable joints (ankles, knees, shoulders, neck), with worst-case ($p95$) discrepancies of 5–9 cm, and only slightly higher under partial occlusion. We show that the larger hip discrepancy (median 4–6.6 cm) is *not* an optical-tracking fault, but a systematic, per-subject difference in how the two systems *define* the hip joint, and we provide evidence of reference-system-related issues (loosely attached sensors, temporal smoothing/step anchoring and step-count-driven pelvis translation on the treadmill), and a camera-calibration contribution, all of which affect the agreement. Methodology and all mathematical details are collected in the Appendices.

1 Introduction and Purpose

Accurate 3D skeleton tracking is a foundation for movement-quality analysis in professional sport: consistent per-joint trajectories let analysts and coaches spot biomechanical deviations (asymmetries, altered joint angles, changes in landing or cutting mechanics) that could be associated with negative impacts on performance or health, and monitor them over time. Sensor suits such as Xsens deliver this today but require instrumenting each athlete and *do not work over large outdoor areas* (no practical coverage of a full pitch and drift/step-integration issues over distance). An **optical, camera-based** system that reconstructs the same skeleton from ordinary video would remove the suit and, crucially, *scale to big open-air spaces* (full-field training and match settings) where the joint-level inertial capture is limited if not completely excluded for various reasons — provided its per-joint positions are trustworthy enough for biomechanical use.

The main purpose of this report is therefore to evaluate the **agreement** between the Optical biomechanics tracking system's (FitWise AI) per-joint 3D skeleton and the reference Xsens 3D skeleton and quantify how closely they match in a controlled environment. We are explicit about a central subtlety: **it is impossible to obtain the ground truth of the joint positions**. The inertial suit does not directly measure joint-center positions; it infers them from IMU signals through a subject-calibrated biomechanical model and is itself subject to sensor and modeling error. Because neither system is a direct measurement of the true joint center, we deliberately speak of the **discrepancy** between the two systems rather than an "error" against a reference. The second subtlety is that the two systems do not share a single joint *definition*, so a raw joint-to-joint distance mixes genuine tracking disagreement with a fixed *definitional* offset (a constant gap present even if both systems tracked perfectly). A large part of this report separates those two contributions and states *which* joints can be compared, *how*, and with *what* residual discrepancy.

2 Setup and Data

2.1 The two systems

Optical biomechanics tracking system (FitWise AI) (hereafter "FitWise AI system"). The system reconstructs a 3D body per detected player and emits a **24-joint** skeleton in **MHR** [1] format. Its joints are defined by the system's internal body model and are placed to reproduce the body surface/pose. The raw estimated skeleton has no absolute size — it is correct in *shape* but not in *scale* due to the setting not containing usually present field markings or other visual cues of known dimensions — so before comparing we rescale it to match the reference (a single size factor shared across the whole clip; Appendix B).

Reference system (Xsens). The baseline reference is an **Xsens** (Xsens/MVN are trademarks of Movella/Xsens Technologies B.V.) inertial mocap suit [2] providing a **23-segment** skeleton at 60 Hz. Xsens is metric and estimates *anatomical* joint centers (e.g. the hip joint center at the femoral-head center of rotation) using its internal biomechanical model driven by the per-subject calibration (entered body dimensions) and IMU fusion; anatomical hip-center placement from pelvis landmarks follows regression relationships such as Harrington et al. [3] and Bell et al. [4]. The Xsens data was post-processed (Sec. 4). We use it as a well-established baseline reference, but it is *not* a direct ground-truth measurement of joint centers: the positions are inferred from IMU signals through the subject-calibrated model and carry their own sensor and modeling error (Sec. 5).

2.2 Physical activities

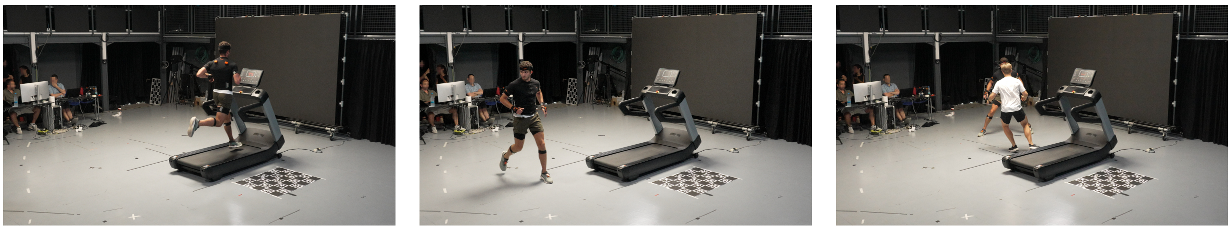
Epidemiological data consistently attribute a large share of non-contact injuries to **high-speed running/sprinting** and to **rapid changes of direction and load** (accelerations, decelerations, cutting). The three activities target exactly this envelope:

- **Treadmill running** exercises the high-speed, straight-line running regime. It is a deliberate compromise: to reach realistic sprint speeds *in the lab setting*, a treadmill is used, because the inertial suit does not support large open-air capture areas.
- **Shuttle runs** exercise repeated accelerate–decelerate–turn cycles (including zig-zag), i.e. the rapid direction- and load-change movements most associated with injury.
- **Partial-occlusion drill** is included specifically to study how stable the *optical* system remains when the tracked athlete is intermittently hidden — a routine occurrence in team sports (contesting for the ball, players crossing the line of sight) and a failure mode that does not affect a body-worn inertial suit.

In addition, the aforementioned activities could stress-test the system approximating movement most relevant to biomechanical analysis applicable in European football.

2.3 Capture environment and data properties

- **Environment:** an indoor lab facility under artificial illumination, with an observable capture area of roughly 32 m^2.
- **Subjects:** 9 athletes (5 male, 4 female), each recorded performing all three activities while wearing the Xsens suit and filmed by a calibrated camera.
- **Camera (geometry & calibration):** a single static camera recording at **4K** (3840 × 2160 px), focal length ≈ 4242 px, mounted ∼ 2.4 m above the floor. The athlete is viewed from a horizontal range of roughly 6–16 m, i.e. a near-frontal/oblique full-body framing throughout. The optical system places the 3D skeleton in world coordinates using this calibration, so any residual reprojection error (sub-pixel for intrinsics; a few pixels for the static-board pose) propagates as a small, roughly constant geometric offset in the optical positions; details and achieved reprojection levels are in Appendix E.
- **Activities:** treadmill running, shuttle runs, and a partial-occlusion drill; see Sec. 2.2 for the rationale and Fig. 1 for example footage. Higher discrepancy is expected under occlusion.
- **Optical-derived stream** (119.88 fps, scaled units) and **Xsens stream** (60 Hz, metric). Each athlete's Xsens export is one continuous session recording (10 047–13 859 samples, ≈ 167–231 s) spanning all activities; each analysed clip is a sub-window of that session. Because the two streams run on different clocks and rates, we line up Xsens onto the FitWise AI system's timeline and, for each video frame, read the Xsens position by blending the two nearest 60 Hz samples (the exact time-mapping is in Appendix A); no smoothing or decimation is applied. Table 1 gives, per athlete and activity, the analysed duration of each clip (the compared video window); the underlying FitWise AI system/Xsens sample counts are retained in the source as a commented table.



(a) Treadmill run (b) Shuttle run (c) Partial-occlusion drill

Figure 1: Example footage of the three activities (athlete A09), captured in the indoor lab: (a) a treadmill run, (b) a shuttle run on the floor, and (c) the partial-occlusion drill, where a second person moves in front of the tracked athlete (imitating contesting the ball). The athlete wears the Xsens suit; the ChArUco board on the floor is used for camera calibration (Appendix E).

Table 1: Per-athlete analysed duration (seconds) of the compared window, per activity. Durations are the compared video frames divided by the video frame rate (119.88 fps); the occlusion drill is shorter than the running activities. The underlying FitWise AI system/Xsens sample counts are in Appendix F.

Athlete	Sex	**Treadmill** (s)	**Shuttle run** (s)	**Occlusion** (s)
A01	M	89.3	36.1	25.8
A02	F	91.7	33.3	13.6
A03	M	112.6	46.7	10.2
A04	M	70.9	17.5	67.1
A05	F	96.3	50.0	12.9
A06	F	101.8	32.3	5.7
A07	M	80.5	60.7	13.5
A08	F	83.8	53.8	12.1
A09	M	53.6	54.4	14.8
Total		780.4	384.8	175.6

3 Agreement Assessment Implementation Details

We evaluate the two systems' agreement in three straightforward steps: align them *temporally*, align them *spatially*, and then measure the residual per-joint distance. Each step is summarized below, while the exact mathematical formulations are provided in the appendices.

3.1 Step 1 — temporal alignment

The two systems run on independent clocks and different frame rates, so a frame number in one does not correspond to the same one in the other. We recover the time offset automatically from the motion itself (matching the moments when the body moves the same way in both streams) and refine it together with the spatial step below. A few clips with an unreliable automatic match were aligned manually (Sec. 5). The full procedure is described in Appendix A. Appendix A also quantifies how the discrepancies respond to the sync — a full-frame timing perturbation moves the distal discrepancy insignificantly — so the sync sensitivity is bounded and small relative to the reported discrepancies.

3.2 Step 2 — spatial alignment

The Optical biomechanics tracking system (FitWise AI) skeleton has the right shape but not the absolute size, and the Xsens body drifts around in the room in certain scenarios (see Sec. 5). So, before evaluating, we (i) rescale the FitWise AI biomechanics tracking skeleton by a single size factor shared across the whole clip, and (ii) rotate and shift it, *separately in each frame*, to best overlay the reference. This is a standard pre-processing step in 3D pose evaluation — specifically, the **PA-MPJPE** protocol used across the human-pose literature [6, 5] was followed. The detailed description as well as the justification of this step can be found in Appendix B.

3.3 Step 3 — joint selection and evaluation

After the two skeletons are aligned, the discrepancy for a joint is simply the distance between the FitWise AI system's joint and its corresponding Xsens point, per frame. The pairs we use are the ankles (`L/R_Ankle`↔Foot), knees (`L/R_Knee`↔Lower Leg), hips (`L/R_Hip`↔Upper Leg), shoulders (`L/R_Shoulder`↔Upper Arm), and a neck term (`Neck`↔midpoint of the two Xsens shoulders).

Comparing two different body models. The FitWise AI system's body model and the Xsens biomechanical model are *different rigs*; therefore, corresponding named joints may not be located at exactly the same anatomical positions, even in the case of a flawless reconstruction. Sec. 6 discusses the correctness of comparing two different body rig models and establishes which joints are safe to compare.

4 Measured Discrepancies Between Optical and IMU-Based Systems

Table 2 reports the per-joint distance discrepancy between the two systems for all three activities, restricted to the joints the two systems define compatibly (ankles, knees, shoulders, neck). The hips are reported separately in Sec. 6, where we explain which joints are valid to compare and why a hip-to-hip distance measures a definitional offset instead. Values are in centimetres. Because the most informative comparison is between *subjects* (Sec. 6), we report the **between-subject mean ± SD**: the mean of the nine per-subject discrepancy levels and their variability across the athletes. This provides a representative measure of the uncertainty expected for a "typical" subject. We also report the median and the worst-case $p95$ to characterize the overall distribution. We deliberately do *not* share a confidence interval computed from individual frames as the primary uncertainty measure. At 120 fps, successive frames are highly correlated and therefore not statistically independent; consequently, a frame-level confidence interval would be misleadingly narrow. Moreover, such an interval would quantify the precision with which the mean discrepancy is estimated rather than the variability of the discrepancy across individual frames or subjects. The reasoning and supporting formulas can be found in Appendix C.

Table 2: Per-joint positional discrepancy (cm) between the FitWise AI system and the Xsens reference, for the joints whose definitions are compatible between the two systems. The headline is the **between-subject mean ± SD** (average of the per-subject discrepancy levels and their spread across subjects); n is the number of subjects contributing (after the per-subject exclusions of Sec. 5). "med" is the median and "$p95$" the worst-case level, summarizing the pooled distribution. Left/right reported separately. The **occlusion** columns come from the partial-occlusion drill (a contesting person intermittently hides the athlete); its discrepancies run slightly higher, as expected. The **hips are deliberately not listed**: the two systems define the hip joint differently, so their distance is a definitional offset rather than a tracking discrepancy — see Sec. 6, which also states which joints are valid to compare and why. The (deliberately not shared) frame-level CI is discussed in the footnote and Appendix C.

	Treadmill			**Shuttle run**			**Occlusion**		
Joint	mean ± SD (n)	med	$p95$	mean ± SD (n)	med	$p95$	mean ± SD (n)	med	$p95$
Ankle (L)	5.4 ± 1.0 (9)	5.1	8.8	5.2 ± 0.9 (9)	5.0	9.2	5.4 ± 1.6 (9)	4.7	9.2
Ankle (R)	5.1 ± 1.0 (9)	5.0	8.3	5.2 ± 1.0 (9)	5.2	9.2	5.8 ± 1.9 (9)	5.3	10.6
Knee (L)	3.3 ± 0.4 (9)	3.1	5.8	3.2 ± 0.5 (9)	3.1	6.1	3.7 ± 1.1 (9)	3.2	6.8
Knee (R)	3.0 ± 0.5 (9)	2.9	5.0	3.2 ± 0.6 (9)	3.2	6.0	3.7 ± 1.0 (9)	3.2	6.3
Shoulder (L)	4.2 ± 1.0 (9)	4.0	7.4	3.1 ± 0.7 (7)	3.0	5.1	2.8 ± 0.6 (7)	2.5	4.2
Shoulder (R)	2.5 ± 0.9 (9)	2.3	4.8	2.5 ± 0.4 (7)	2.5	4.2	2.6 ± 0.4 (7)	2.6	4.7
Neck (mid-shoulders)	2.1 ± 0.4 (9)	2.1	3.0	2.1 ± 0.5 (9)	2.0	3.6	2.1 ± 0.5 (9)	2.1	3.3

Hips are excluded from this table by design (Sec. 6); their levels are reported separately in Table 3 as a definitional offset. The shoulder exclusions of Sec. 5 (A03 left, A06 both, A07 right) apply to the shuttle and occlusion sets, hence n=7 for the shoulders there. See Appendix C for a detailed justification of the non-usage of a frame-level confidence interval.

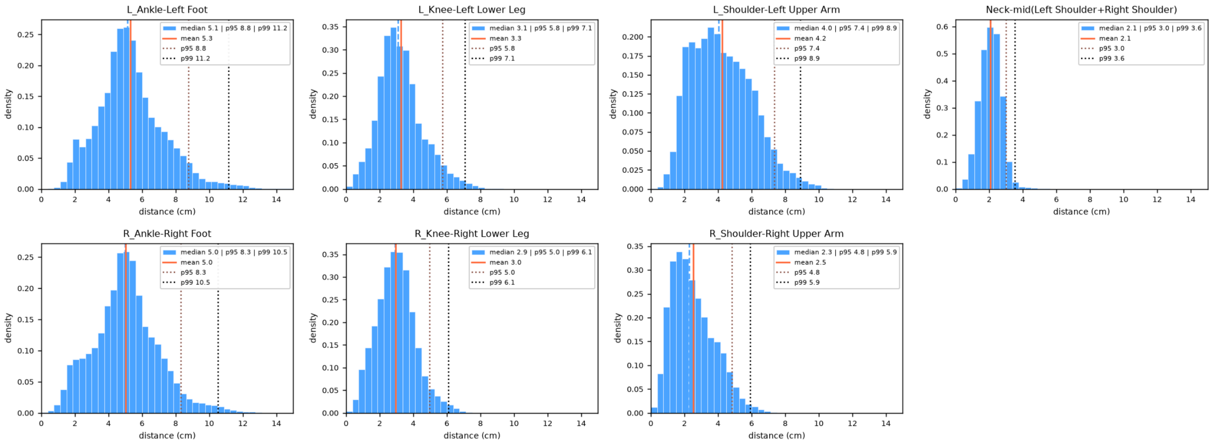


Figure 2: Treadmill: per-joint distance-discrepancy histograms pooled across all athletes (single distribution per joint; density-normalized, shared axes). See Sec. 6 on joint selection.

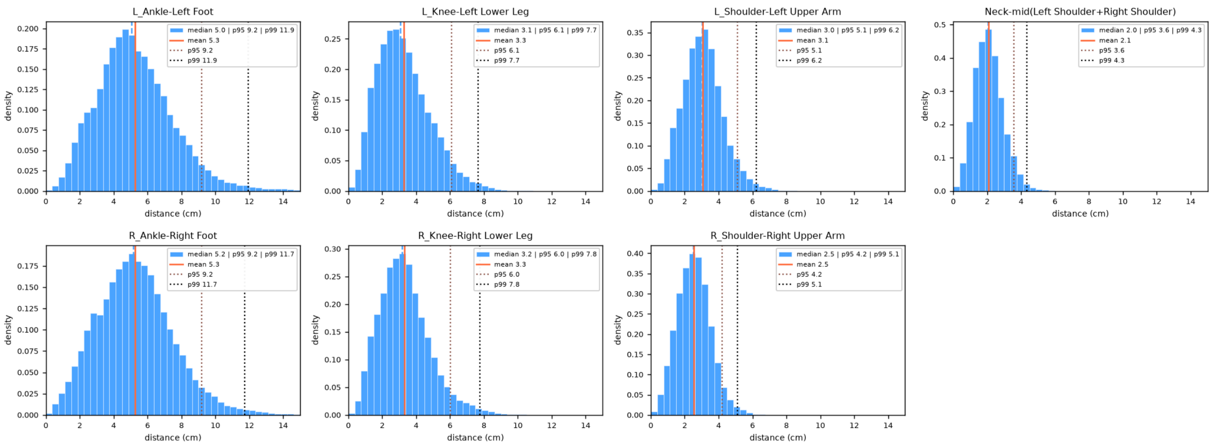


Figure 3: Shuttle run: per-joint distance-discrepancy histograms pooled across all athletes (single distribution per joint; density-normalized, shared axes).

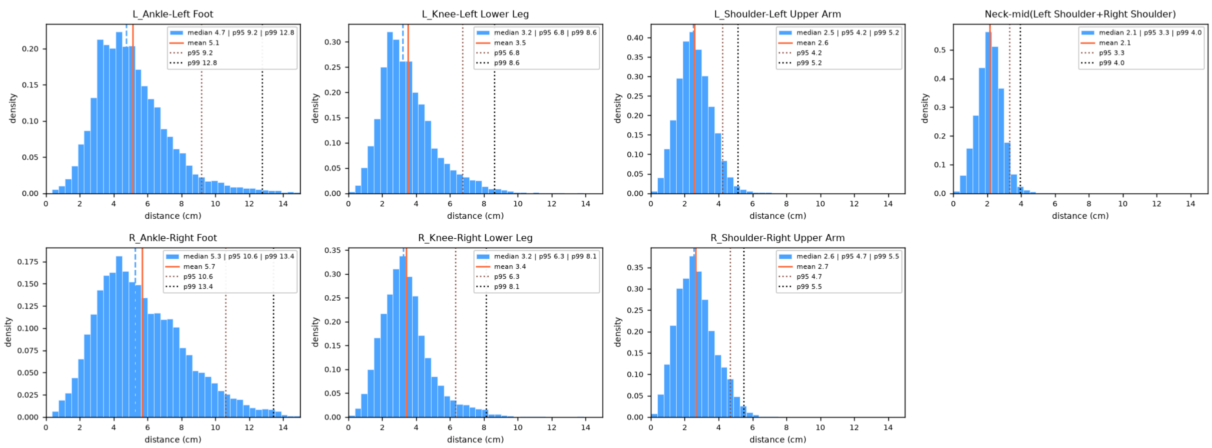


Figure 4: Partial-occlusion drill: per-joint distance-discrepancy histograms pooled across all athletes (single distribution per joint; density-normalized, shared axes). A contesting person intermittently hides parts of the tracked athlete, giving slightly heavier tails than the running activities (compare Figs. 2, 3).

Discrepancy-over-time behaviour. Plotting the per-joint discrepancy against clip progress (Fig. 5) shows that on the treadmill the distal joints (ankles, knees) drift *upward* as running speed increases, then relax when the speed drops — consistent with velocity-dependent lag/smoothing. We quantify this per clip by (a) how much the distal discrepancy (ankles+knees) grows from the start to the end of the clip, and (b) whether that growth is a steady trend or just noise (a rank-correlation "trend strength" between 0 and 1; details in Appendix C). On the treadmill the discrepancy grows by $+\mathbf{0.97} \pm \mathbf{0.85}$ cm over a clip (mean ± SD across the 9 clips; statistically significant, p=0.009), and the growth is a *steady* trend with progress (trend strength +0.28), as expected when the belt speed ramps up. In the shuttle runs the growth is similar in size ($+0.64 \pm 0.78$ cm per clip) but *not* steady: the trend strength is near zero (+0.11, less than half the treadmill value and driven by one clip), i.e. the discrepancy does *not* rise systematically with elapsed time. Instead it appears in bursts aligned with the accelerate/turn cycles (Fig. 6), confirming the discrepancy tracks instantaneous joint speed rather than time-in-clip; the shuttle run, having no monotonic speed ramp, shows no monotonic discrepancy growth. Figs. 5 and 6 show a single representative athlete (A06) so the shape of the trend stays legible; the growth and trend-strength figures quoted above are computed over all nine athletes. Both figures show the *same deliberately selected subset* of joints, chosen because they span a range of segment speeds: as the run gets faster the **ankles** speed up the most (the fastest-moving segment in a stride), the **knees** also speed up but less, and the **neck**/torso center keeps a roughly constant speed throughout. That makes this example a clean visual test of the velocity-dependence claim: the discrepancy grows exactly for the joints whose speed grows, and stays flat for the joint whose speed does not. Restricting both figures to this same subset also lets the two activities be read panel by panel.

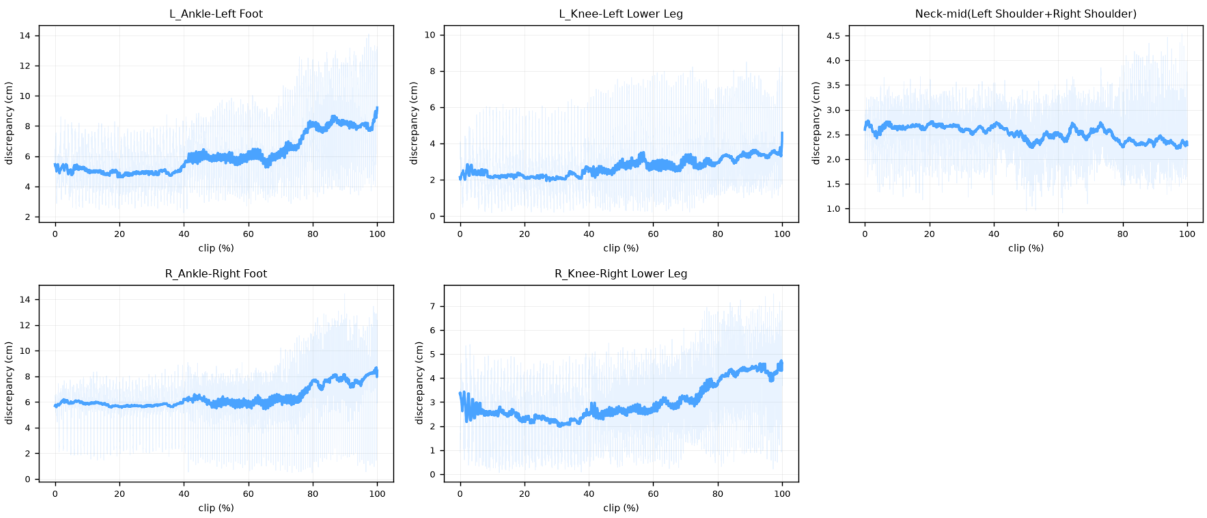


Figure 5: Treadmill: per-joint discrepancy trend (bold rolling median over the faint per-frame trace) vs. clip progress, for a single representative athlete (A06) to keep the trend readable. A selected subset of joints spanning a range of segment speeds: as the treadmill speed ramps up the **ankles** accelerate the most and show the largest growth, the **knees** accelerate less and grow correspondingly less, while the **neck**/torso center holds a roughly constant speed and stays flat — discrepancy growth follows joint speed, not elapsed time.

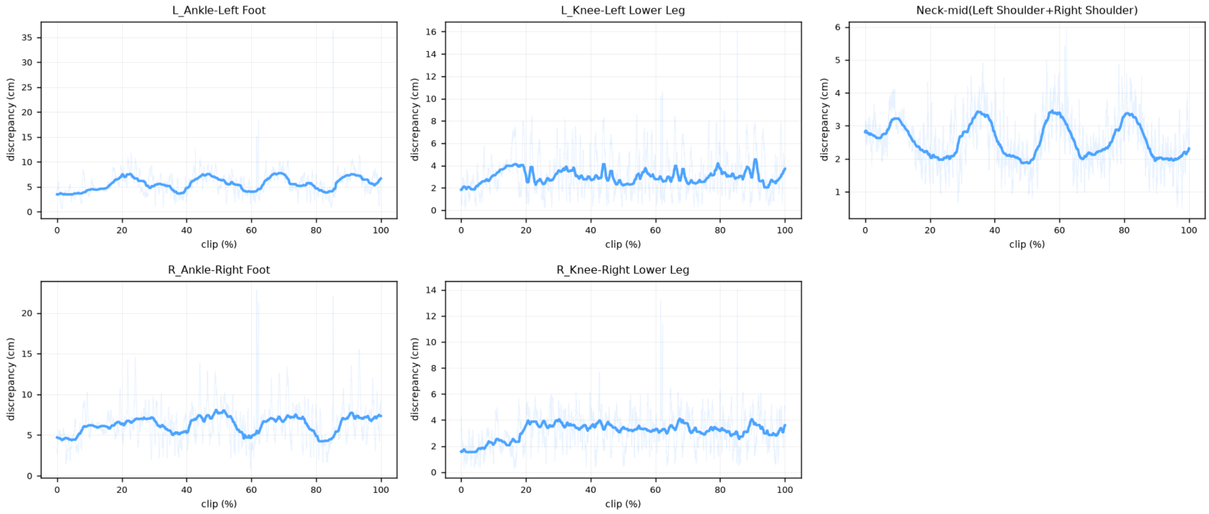


Figure 6: Shuttle run: per-joint discrepancy trend (bold rolling median over the faint per-frame trace) vs. clip progress, same athlete and same joint set as Fig. 5. Discrepancy appears in repeated bursts locked to the accelerate/decelerate/turn cycles rather than trending with time — contrast the steady climb in Fig. 5. The quoted statistics are computed across all nine athletes.

On Xsens data post-processing. Our exports come from MVN's offline *reprocessing* pass, which fits the trajectory to detected foot strikes and a static-ground contact assumption. Fig. 7 shows the consequence — the reported foot speed collapses from $\approx 5\,\mathrm{m/s}$ mid-swing to $\approx 0.3\,\mathrm{m/s}$ at every ground contact, even though the foot is riding a treadmill belt moving backwards at running speed and therefore cannot be near standstill. Consistent with this, the Xsens ankle acceleration is compressed $\sim 5-7\times$ relative to the FitWise AI system in **all 9 subjects** (ratio $\mathbf{5.6 \pm 0.6}$; Appendix D). The evidence is consistent with the presence of temporal smoothing and step-count-based post-processing in the reference skeleton. Both factors *may cause a lag behind the athlete's actual joint motion*, and the faster the joint moves, the further it lags — which may explain the velocity-dependent growth seen above. Part of the distal discrepancy may therefore be attributable to reference-system lag rather than the FitWise AI system's inaccuracy.

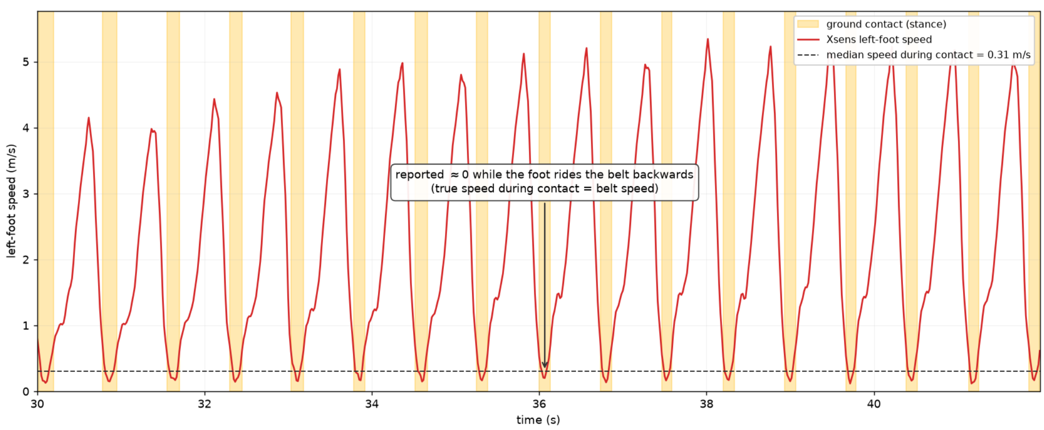


Figure 7: Xsens left-foot speed over 12 s of treadmill running (athlete A01), with the detected ground-contact phases shaded. Speed drops from $\approx 5\,\mathrm{m/s}$ mid-swing to a median of $0.31\,\mathrm{m/s}$ during every contact (dashed line). A near-zero planted foot is what a static-ground (zero-velocity) contact assumption produces — but on a treadmill the foot is riding a belt moving backwards at the running speed, so its true speed during contact is the belt speed. The plateaus are an imposed contact model, not a measurement.

5 Xsens Reference Data Considerations

The Xsens stream is treated as the reference but, like any measurement system, is subject to measurement uncertainty. The collected evidence showed that the following factors may affect the comparison and motivate specific exclusions.

- **Loosely fixed sensors corrupting segments.** On some subjects a poorly attached IMU makes its own segment estimate unreliable, producing a large one-sided discrepancy unrelated to the FitWise AI system. Importantly, the effect is *not* confined to the loose segment: Xsens fuses all IMUs through a single articulated biomechanical model with kinematic constraints and a contact/step prior, so a corrupted segment is partially propagated to neighbouring joints (e.g. a loose upper-arm IMU perturbs the shoulder and, through the shoulder girdle, the neck/torso estimate). We therefore exclude the directly affected joints per subject rather than trusting a local repair: `L_Shoulder` for A03, *both* shoulders for A06, and `R_Shoulder` for A07 (in both the shuttle and occlusion sets); and the treadmill hips for A06.
- **Step-count-driven pelvis translation (treadmill).** Xsens integrates a step counter to place the root in world space. On a treadmill the subject is stationary while the belt moves, so the integrated pelvis *translates* across the (virtual) floor even though the athlete does not advance. Our per-frame translation t_f cancels this drift for the comparison, but it is a reminder that the Xsens *global* position is not a valid world reference on a treadmill; only relative/segment geometry is used here.
- **Temporal regularization / step-anchoring.** The Xsens reference is temporally smoother than the per-frame reconstruction (much lower acceleration, Appendix D), and Fig. 7 shows the mechanism directly: the foot speed is driven to ≈ 0 at every ground contact, i.e. a static-ground contact assumption that a moving treadmill belt violates. This negatively contributes to the apparent distal-joint agreement during fast motion.

6 Validity of the Joint Comparison

The FitWise body model and the Xsens biomechanical model define joints in different ways, so a raw joint-to-joint distance contains two parts: (i) a *definitional offset* — a per-subject, near-constant gap between the two models' definition of that joint, present even with zero tracking disagreement — and (ii) the *tracking discrepancy* we actually want. The space-alignment of Sec. 3.2 removes the global pose/scale difference but *cannot* absorb a joint-specific definitional offset, which stays as a constant bias at that joint. A distance is therefore a valid agreement measure only where the two definitions (nearly) coincide. Below we test this per joint and find it holds for the distal limbs and torso center but fails at the hips.

Hip discrepancy: differences in joint definition rather than tracking error . The pooled hip histogram is *bimodal* (Fig. 8). The second peak is not a heavy tail present in every athlete — it is produced by a *subset of subjects*. The per-subject mean hip discrepancy splits cleanly into a "good" group (A01–A04, A08: 2.4–4.0 cm) and a "bad" group (A05, A06, A07, A09: 4.7–6.6 cm); each subject's mean is estimated to within ±0.1 cm (frame-level 95 % CI), and, more conservatively, the between-subject spread of the two groups does not overlap — so the two groups are distinct rather than a sampling artefact. Three facts show this is a fixed offset and not tracking disagreement:

1. **Activity-invariance.** The same subjects are "bad" in *both* the treadmill and shuttle runs, with almost identical hip discrepancy in the two very different activities (Table 3) — a motion-driven discrepancy would change with the motion.

2. **Joint-locality.** The bad-hip subjects have *normal* knee discrepancy (3.0–3.7 cm, indistinguishable from the good group): the problem is hip-specific, not a globally worse reconstruction for those athletes (Table 3, "knee" column).

3. **Activity-preserved split.** If the second peak were motion noise it should shrink in one activity; instead the good/bad split and the per-subject means are preserved across activities, i.e. the offset travels with the subject, as a definitional bias must.

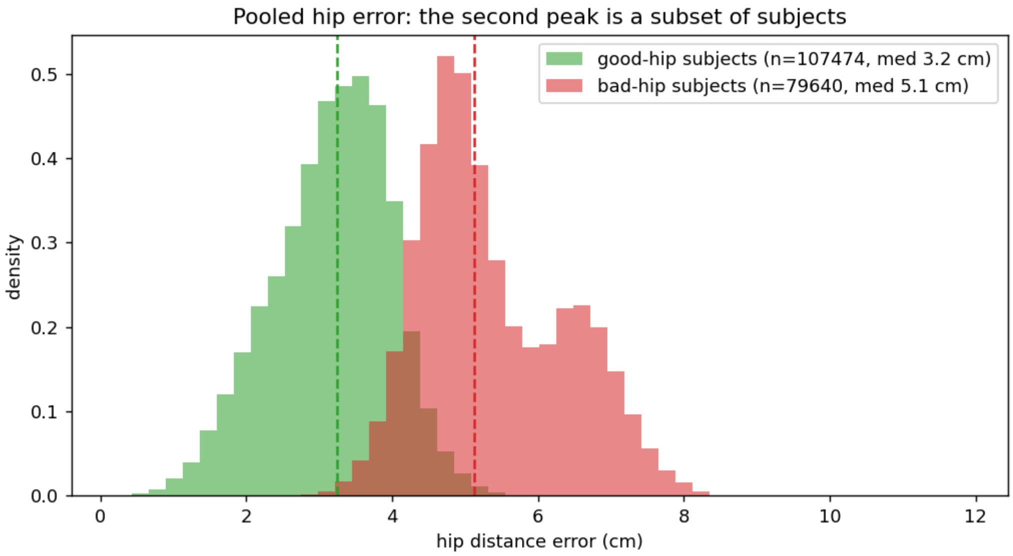


Figure 8: Pooled hip distance discrepancy split by subject group. The "good" and "bad" subject groups occupy the two peaks of the bimodal pooled distribution: the second peak is a per-subject definitional offset, not a within-subject tail.

Table 3: Per-subject mean hip discrepancy (L/R averaged, cm) is stable across activities — a definitional offset, not a motion/tracking disagreement — while the same "bad-hip" subjects (A05, A06, A07, A09) have *normal* knee discrepancy. Hips are reported here rather than in Table 2 for exactly this reason. The treadmill hip column reports the raw value (A06's inflated hip is the loose-sensor subject, Sec. 5).

Subject	Hip (TM)	Hip (SR)	Δ	Knee (TM)
A01	2.8	3.1	+0.3	3.2
A02	2.4	2.9	+0.5	2.5
A03	3.1	4.0	+0.9	3.1
A04	4.0	4.1	+0.2	3.2
A05	5.0	5.3	+0.3	3.0
A06	6.6	6.6	+0.0	3.0
A07	4.7	4.9	+0.2	3.3
A08	3.7	4.1	+0.4	3.2
A09	4.7	4.7	+0.0	3.7

The mechanism is a genuine modeling difference. Xsens places the hip at the *anatomical* hip joint center (femoral-head rotation center), so the left-HJC–pelvis–right-HJC forms a near-straight lateral line through the pelvis. The FitWise AI system's hips are placed by its body model *medial and superior* to the true HJC, forming a shallow "V" from the (higher, central) pelvis root down-and-out to the estimated hips. The gap is a fixed per-subject vector in the pelvis frame; it is a known limitation of comparing a surface/pose-driven body model against an anatomical rig, and is typically addressed by re-mapping the estimated joints to anatomical hip-joint centres [3, 4] or by using an anatomically rigged body model. We therefore treat the hips as carrying a per-subject definitional offset and do *not* report them as a raw tracking agreement; quantifying the residual after removing that offset is out of scope here.

Safe-to-compare joints. Combining the bias analysis (activity-invariance, joint-locality) across both datasets (Table 4):

- **Ankles, knees** — motion-dominated discrepancy that scales with joint speed, no per-subject constant offset: *directly comparable (no definitional offset).*
- **Shoulders** — motion-dominated and comparable; report left/right separately (a real, view-dependent side asymmetry exists) and note A05's L-shoulder shows a mild partial offset.
- **Neck (mid-shoulders)** — smallest and tightest discrepancy: comparable, but it is a *constructed midpoint* (torso-center agreement), so it cannot reveal left/right asymmetry.
- **Hips** — dominated by a stable per-subject definitional offset; *not* valid as raw agreement, only after bias removal.

Table 4: Comparison-validity verdict per joint.

Joint	Discrepancy character	Verdict
Ankles (L/R)	Motion-driven, no offset	Directly comparable (no offset)
Knees (L/R)	Motion-driven, no offset	Directly comparable (no offset)
Shoulders (L/R)	Motion-driven, side-asymmetric	Safe; report L/R separately
Neck (mid-shoulders)	Small, motion-driven	Safe; torso-center, no L/R info
Hips (L/R)	Per-subject constant offset	Bias-corrected only (definitional)

7 Conclusion

For the joints that share a compatible definition between the FitWise AI system and Xsens — **ankles, knees, shoulders, and the neck/torso-center** — the optical-vs-inertial positional discrepancy is:

- **Median** 2–5 **cm** per joint, essentially identical across the treadmill and shuttle-run activities, and only slightly higher under the partial-occlusion drill (Table 2);
- $p95$ of $\approx$ 5–9 cm, with the **distal joints** (ankles) the largest and the **torso/neck** the smallest ($\sim$2 cm);
- **Velocity-dependent**: distal-joint discrepancy grows with running speed (treadmill ramp) and with acceleration bursts (shuttle turns), consistent with estimation/smoothing lag on fast segments.

The apparently large **hip** discrepancy (4–6.6 cm) is **not** an optical-tracking fault: it is a stable, per-subject *definitional* offset between the FitWise AI system's and Xsens's (anatomical HJC) hip conventions, demonstrated by its activity-invariance and joint-locality (the same subjects have normal knee discrepancy). We therefore report hips as a definitional bias and do not treat them as a raw agreement.

Bottom line. The FitWise AI system agrees with the inertial suit on limb and torso joint positions to a **median of a few centimetres**, with the dominant residuals attributable to (a) fast-segment lag at the extremities and (b) reference-side Xsens issues (loose sensors, treadmill step-integration), rather than to a systematic optical bias — except at the hips, where the discrepancy is a coordinate-definition difference rather than a tracking fault. Given that neither system is a direct ground-truth measurement, this level of agreement supports using the FitWise AI system as a practical, suit-free alternative for biomechanical monitoring, including in the large open-air settings where inertial capture is impractical.

A Temporal Synchronization (Details)

The two streams run on independent clocks. We recover the time map automatically by (i) an FFT cross-correlation of a rotation/scale-invariant motion signal (built from the correspondence joints) to seed a shift, then (ii) a joint optimization of an affine time map

$$t_{\text{xsens}} \;=\; a\, t_{\text{pipe}} + b, \qquad t_{\text{pipe}} = f/\text{fps}, \tag{1}$$

(clock scale a constrained near 1 and the offset b) together with the spatial fit of Appendix B, minimizing the pooled correspondence residual. For each video frame, the Xsens position is obtained by *linear interpolation* between the two bracketing 60 Hz samples (per coordinate); no smoothing or decimation is applied. Clips with a low cross-correlation confidence fall back to a manual seed offset; a few clips were re-synced by hand (Sec. 5).

Sync-sensitivity test. Because the time map is fit jointly with the spatial alignment, it could in principle lower the reported discrepancy by trading the offset b against distal motion. We bound this: holding the clock scale a fixed, we re-ran the *same* spatial fit at the fitted b and at $b \pm 1$ video frame ($\pm 1/119.88$ s) on every treadmill clip and measured the change in pooled distal (ankle+knee) discrepancy. The discrepancy moves by a median of only 0.44 cm (max 0.59 cm; $\approx$10% of the $\sim$4.3 cm base) for a full-frame offset, so a frame-level sync error cannot account for the low distal numbers.

B Spatial Alignment and PA-MPJPE (Details)

The FitWise AI system's skeleton is unscaled and the Xsens pelvis drifts, so we fit, in one closed-form pass, a *single global* scale s shared across all frames together with a *per-frame* rotation R_f and translation t_f:

$$x^f \;\approx\; s\, R_f\, p^f + t_f, \tag{2}$$

over a fixed correspondence set. The joints entering the s, R_f, t_f estimate are the **pelvis** plus the paired ankles, knees, hips, shoulders, and the neck/shoulder terms, i.e. the index set {Pelvis, L/R_Ankle, L/R_Knee, L/R_Hip, L/R_Shoulder, Neck} on the FitWise AI system side matched to {Pelvis, L/R Foot, L/R Lower Leg, L/R Upper Leg, L/R Upper Arm, L/R Shoulder} on the Xsens side. Because the centred Procrustes [5] rotation is scale-independent, each R_f solves in closed form, then the pooled global scale is

$$s \;=\; \frac{\sum_f \langle R_f\, \tilde{p}^f,\, \tilde{x}^f \rangle}{\sum_f \|\tilde{p}^f\|^2}, \tag{3}$$

and finally t_f (with $\tilde{\cdot}$ the centred coordinates). The global s recovers a consistent metric size while R_f, t_f absorb orientation and the Xsens pelvis drift. The reported per-joint discrepancy is the Euclidean distance between the aligned FitWise AI system joint and its corresponding Xsens segment origin, per frame.

Why the alignment does not flatter the result. s is a *single scalar shared across every frame* (it cannot bend an individual pose), and R_f, t_f are a rigid motion which *preserves all inter-joint distances and angles.* Any deformation — a mis-placed limb, a wrong joint angle, a definitional offset — survives the transform untouched and shows up in the residual. What is removed is only the rigid pose and the one global size, i.e. exactly the world-placement we are *not* measuring.

Relation to MPJPE / PA-MPJPE. A single global scale plus a per-frame rigid (rotation+translation) Procrustes alignment before measuring mean per-joint position error is exactly **PA-MPJPE** (Procrustes-aligned mean per-joint position error), the standard reconstruction metric popularized by Human3.6M [6]. We use the same computation, but because our reference is another estimating system rather than a direct ground truth, we report the resulting quantity as a per-joint *discrepancy*; it is numerically comparable to PA-MPJPE numbers in the 3D human-pose literature, with the caveat that our reference is a different skeleton model (Sec. 6).

C Statistical Treatment (Details)

Reported uncertainty. We headline the *between-subject* mean ± SD (n=9 per-subject means), which reflects how much a typical subject's discrepancy level varies across athletes.

Why not a frame-level CI. A pooled 95 % CI of the mean, even with an autocorrelation-corrected effective sample size

$$N_{\text{eff}} \;=\; N\,\frac{1-\rho_1}{1+\rho_1}, \tag{4}$$

comes out at the sub-mm level (e.g. ankle $\approx\pm 0.1$ cm) — far tighter than the histograms suggest. This is expected: the CI of the *mean* shrinks like $\text{SD}/\sqrt{N_{\text{eff}}}$ and measures how well we know the *average bias*, not how much a single frame deviates (the distribution is genuinely wide — see the $p95$ column). Moreover, a *lag-1* AR(1) correction *understates* the true autocorrelation of a 120 fps kinematic signal, whose correlation persists over many frames (an integrated-autocorrelation-time or stride-block bootstrap would give a smaller N_{eff} and a wider interval); we therefore do not lean on this CI at all and use the between-subject SD (n=9) as the reported uncertainty. The genuine spread of per-frame discrepancies is instead read off the $p95$ column of Table 2 and the histograms (Figs. 2, 3).

Discrepancy-over-time trend statistics. Per clip we fit a linear regression of the mean distal-joint discrepancy (ankles+knees) against normalized clip progress (0–100%) to get the growth (cm per clip), and measure the Spearman rank correlation ρ_s of discrepancy vs. progress as a monotonic-trend check (the "trend strength" quoted in the main text). Significance of the per-clip growth is a one-sample t-test across the 9 clips (treadmill: t_8=3.44, p=0.009; mean ρ_s=+0.28). Shuttle: ρ_s=+0.11.

Per-subject means. Each subject's per-joint mean is estimated to within ±0.1 cm (frame-level 95 % CI); the good/bad hip groups are separated by more than their between-subject spread, so the split is not a sampling artefact.

D Xsens Smoothing Metric (Details)

We compare ankle/foot dynamics in each rig's own metric frame, normalized to body-lengths so the two are unit-comparable. The smoothness comparison uses the *acceleration-compression ratio*

$$\rho_a \;=\; \frac{p95(|a|_{\text{FitWise AI system}})}{p95(|a|_{\text{Xsens}})}, \tag{5}$$

i.e. the ratio of the 95 th percentile absolute acceleration of the FitWise AI system to that of Xsens; $\rho_a > 1$ means the reference is smoother. Across the 9 treadmill subjects $\rho_a = 5.6 \pm 0.6$ (median 5.4, range 4.9–7.3).

Stance detection (Fig. 7). The shaded ground-contact phases are found from the Xsens foot-speed trace alone: a sample is in stance when its speed falls below 15 % of the swing peak (the $p95$ of the speed in the plotted window), and a run of such samples counts as a contact only if it lasts at least 80 ms (which rejects single-sample dips between strides). In the window shown this yields 17 contacts covering 22 % of the window, with a swing peak of 4.87 m/s against a stance threshold of 0.73 m/s and a median in-contact speed of 0.31 m/s. Speeds here are plain metric ($\mathrm{m\,s^{-1}}$), not body-length-normalized, since only one modality is plotted.

E Camera Calibration (Details)

The FitWise AI system needs a metric camera model to place the 3D skeleton in world coordinates, so a dedicated calibration video (a scene filmed once with the static camera) is processed with two ChArUco boards:

- a **moving** board (8×5 squares, `DICT_4X4`) waved in front of the camera to span the image and depth range — used to estimate the **intrinsics** (focal lengths, principal point, lens distortion); and
- a **static** board lying on the floor (5×7 squares, `DICT_5X5`) that defines the **world coordinate system**; the fixed camera pose (**extrinsics**) is expressed relative to it.

Each calibration frame is processed in two steps for accuracy: the board is first *detected* on a down-scaled copy (fast, robust), then its chessboard corners are *refined* on the full-resolution frame with sub-pixel corner refinement.

Intrinsics. Corner correspondences from all usable frames are fit with `cv2.calibrateCamera` (rational distortion model), the worst 20 % of frames by reprojection residual are dropped, and the model is re-fit on the survivors. On this capture (3840×2160, 286 calibration frames) the reprojection RMS improves from **0.48 px** (first pass) to **0.39 px** on the retained 229 frames — i.e. *sub-pixel* intrinsics. The recovered focal length is $\approx$4242 px with the principal point near the image center.

Extrinsics. For every frame that sees the static floor board we solve `cv2.solvePnP` with those intrinsics to get the board pose in the camera frame, then average the poses (translation arithmetic mean, rotation averaged on $SO(3)$). The mean PnP reprojection residual of the static board is **2.73 px** over 225 frames.

Effect on the comparison. These residuals are small but non-zero: the sub-pixel intrinsics and the few-pixel static-board pose translate into a small, roughly constant geometric offset in the optical joint positions (a few millimetres at the working range, growing mildly with distance, Sec. 6). Calibration is therefore one contributor to the reported discrepancy, alongside the reference-side Xsens issues; it is not a ground-truth measurement of the camera geometry either.

F Data and Sample Counts (Details)

Table 1 in the main text reports analysed *durations*; Table 5 below gives the underlying *sample counts*. For each activity, "FitWise AI" is the number of video frames actually compared (inside the synchronized window, Sec. 5) and "Xs. used" is the number of native 60 Hz Xsens samples spanning that same window (used *before* interpolation). Because the camera runs at $\approx 2\times$ the Xsens rate, "Xs. used" is about half of "pipe". "Xs. raw" is the whole continuous session export (all activities plus idle/setup) from which each clip's window is drawn. Durations in Table 1 are "pipe" divided by the video frame rate (119.88 fps).

Table 5: Per-athlete sample counts underlying the durations of Table 1. "FitWise AI" = video frames compared; "Xs. used" = native 60 Hz Xsens samples in that window; "Xs. raw" = whole session export.

		Treadmill		Shuttle run		Occlusion		
Athlete	Sex	FitWise AI	Xs. used	FitWise AI	Xs. used	FitWise AI	Xs. used	Xs. raw
A01	M	10 703	5 357	4 322	2 163	3 095	1 549	11 117
A02	F	10 990	5 500	3 992	1 998	1 632	817	12 504
A03	M	13 498	6 756	5 600	2 803	1 222	612	11 716
A04	M	8 500	4 254	2 092	1 047	8 040	4 024	10 047
A05	F	11 546	5 779	5 997	3 001	1 541	771	12 064
A06	F	12 204	6 108	3 871	1 937	681	341	13 859
A07	M	9 645	4 827	7 273	3 640	1 623	812	10 869
A08	F	10 046	5 028	6 455	3 231	1 446	724	11 877
A09	M	6 425	3 216	6 525	3 266	1 776	889	11 389
Total		93 557	46 825	46 127	23 086	21 056	10 539	—

References

[1] A. Ferguson et al. MHR: Momentum Human Rig. *arXiv preprint arXiv:2511.15586*, 2025. Available: https://arxiv.org/abs/2511.15586.

[2] D. Roetenberg, H. Luinge, and P. Slycke. Xsens MVN: Full 6DOF human motion tracking using miniature inertial sensors. *Xsens Technologies Technical Report*, 2013.

[3] M. E. Harrington, A. B. Zavatsky, S. E. M. Lawson, Z. Yuan, and T. N. Theologis. Prediction of the hip joint center in adults, children, and patients with cerebral palsy based on magnetic resonance imaging. *Journal of Biomechanics*, 40(3):595–602, 2007.

[4] A. L. Bell, D. R. Pedersen, and R. A. Brand. A comparison of the accuracy of several hip center location prediction methods. *Journal of Biomechanics*, 23(6):617–621, 1990.

[5] J. C. Gower. Generalized Procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

[6] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI)*, 36(7):1325–1339, 2014.