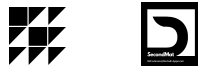




Tracer: An AI-Powered Boxing Training Platform

Glove sensing, per-punch quality scoring and real-time
coaching on consumer hardware

Technical whitepaper prepared for Fierce Reflex · October 2026
Delivered by CSZ in collaboration with SecondMat



Abstract

Tracer is a boxing training platform: it scores **how well** each punch is thrown rather than only which punch it was, coaches the athlete during the session, and tracks progress across sessions from an app. Two glove-worn inertial boards and a phone camera feed a segmentation and scoring pipeline whose output is a structured, coach-defined rubric; commentary is generated from those scores, never end to end from raw signals, so the rubric remains the interpretable layer between perception and language. This paper reports what has been measured on hardware, specifies the custom glove board, sets out the scoring and annotation design, and gives manufacturing cost estimates. On held-out data the current wrist classifier reaches 0.693 three-way punch-type accuracy across every captured punch against 0.333 chance, and 0.887 when the two arc punches are merged; the camera path reaches 0.614 mean-class recall held out by subject. Two measurements govern the hardware specification: 8.0 per cent of real punches exceed the development accelerometer's ± 16 g range, while angular rate peaks at 1776 $^{\circ}/s$, comfortably inside the ± 2000 $^{\circ}/s$ limit — so our own data justifies the high-g channel directly and sets the gyro specification with headroom. Results to date come from a single athlete, and multi-athlete validation is the next milestone. Manufacturing cost at roughly 1,000 units is estimated at **$\approx \$24.50$ per glove**, or **$\approx \$49$ per pair**, falling to **$\approx \$17$ per glove** if a single combined sensor verifies. All costs are October 2026 estimates of cost to manufacture only.

1 The problem

A coach cannot watch fifteen to thirty athletes closely at once. Each sees a few corrected punches an hour; the rest are thrown unobserved, and a fault repeated unobserved is a fault being trained in. The constraint is not coaching knowledge but coaching attention.

Existing products count punches and report speed. Counting is the easy half and the less useful half: an athlete does not improve by learning they threw 312 punches, but by learning that their lead hand drops on the third jab of every combination. Recent work is explicit on this point. SPAR^[4] reports coaches saying a bare score is useless without explanation, and that their binding constraint is how many athletes they can watch; the CoachMe^[5] boxing evaluator adds that concise instructions beat detailed ones, where general language models tend to over-explain.

Tracer therefore targets **per-punch quality on a coach-defined rubric**, with feedback timed to how motor learning actually works, and an explicit structured layer between what the sensors see and what the athlete hears.

2 System overview

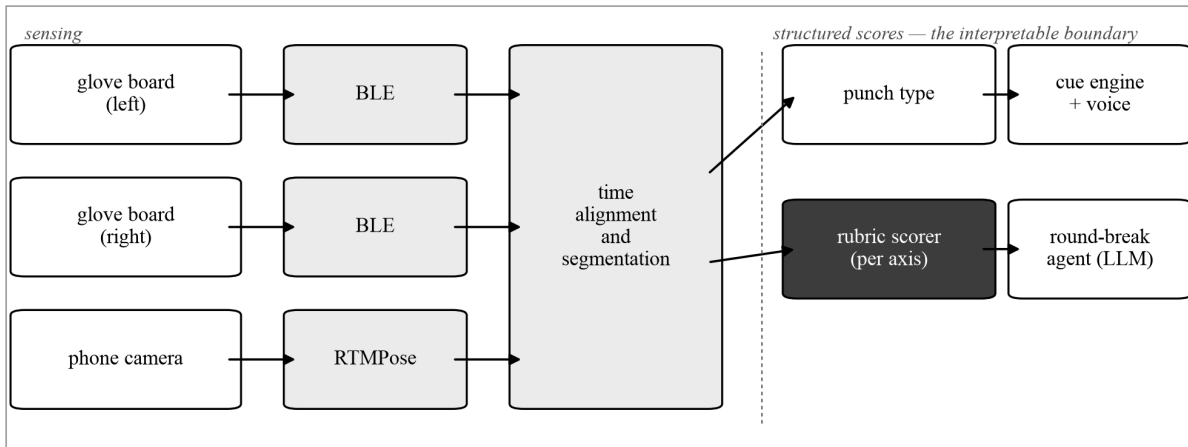


Figure 1. System architecture. Two glove boards and the phone camera are time-aligned and segmented into punch windows; a transformer emits punch type and a score per rubric axis. Everything to the right of the dashed line consumes **structured scores**, not raw signal — the cue engine and the round-break agent never see sensor data. That boundary is what makes the system auditable: any spoken cue can be traced to the axis score that triggered it.

Inputs are two glove-mounted inertial boards over Bluetooth and phone-camera pose estimation via RTMPose^[8]. The pipeline runs punch detection → punch-type classification → rubric scoring → cue engine → voice, with a language-model agent at round breaks and session end.

The design commitment worth stating plainly is that **commentary is generated from structured scores and never end to end from raw signals**. An end-to-end model that speaks directly from accelerometer traces cannot be audited, cannot be corrected axis by axis, and fails silently when it drifts. A rubric can be checked against a coach, disagreement can be localised to an axis, and a cue can be withheld when the relevant score is uncertain.

3 What has been measured

Everything in this section ran on hardware. Punch-type classification is a prerequisite for quality scoring — a rubric is per punch type — so it was built and evaluated first.

actual class	straight	60 74%	12 15%	9 11%
	hook	4 5%	54 68%	22 28%
	uppercut	2 3%	24 31%	51 66%
		straight	hook	uppercut
		predicted class		

Figure 2. Wrist punch-type classifier over all 238 captured punches, pooled across both stance folds, so every clip is scored by a model that never saw its stance. Accuracy and mean-class recall are both 0.693 against 0.333 chance. The boxed cells are the residual error: hook and uppercut exchange 22 and 24 clips, almost symmetrically. *idle* is excluded — it is a runtime class letting the model decline a window, not a punch type, and a 5-clip row beside three 70-clip rows invites a meaningless average. Restricting to punches above the 4.0 g detection threshold — all the live system ever classifies — gives 0.754 over 224 clips. The full corpus is shown here so the figure describes the widest set of inputs rather than the narrowest.

Held out by stance — a model trained on orthodox tested on southpaw and the reverse — the wrist classifier reaches **0.693** across straight, hook and uppercut, over every punch captured. Merging the two arc punches into a single class gives **0.887**; straights then separate into jab and cross with no model at all, because the sensor reports which wrist fired and stance is a setting. The camera path, held out by subject on 5,451 public clips, reaches 0.663 top-1 and 0.614 mean-class recall, with 176 of 298 uppercuts returned as hooks.

Both tiers meet their limit at the same boundary, for the same physical reason: a monocular camera discards the vertical component that defines an uppercut, and the wrists cannot observe the trunk rotation that separates the two arcs. This is the single strongest argument for camera-IMU fusion (§8) and it is a measurement, not a conjecture.

Three engineering problems were identified and solved during this work. Each would have degraded a shipped product silently, and each is recorded because the lesson generalises. **A sensor-frame mismatch** — a model fitted before per-board calibration existed, then fed rotated input — silently returned hook for 62 per cent of all punches while validating at 0.74. A rotated input is not noise: a model does not report that it has not seen something, it extrapolates confidently into whichever class occupies the largest region of feature space. The fix was to use only rotation-invariant features, verified over 274 clips against 25 random rotations with zero predicted labels changed. **A detection threshold** set at 3.0 g rejected none of the non-punch clips in its own validation set, and could not have: every clip was recorded by that same detector, so by construction nothing in it could fail that test; the threshold was reset against the measured distribution instead. **Radio throughput** was measured at 33 Hz against a declared 104 Hz, identifying the Bluetooth batching as the largest available gain in signal quality — and one that costs no new data to collect.

4 Hardware: the custom glove board

4.1 The sensing requirement, measured

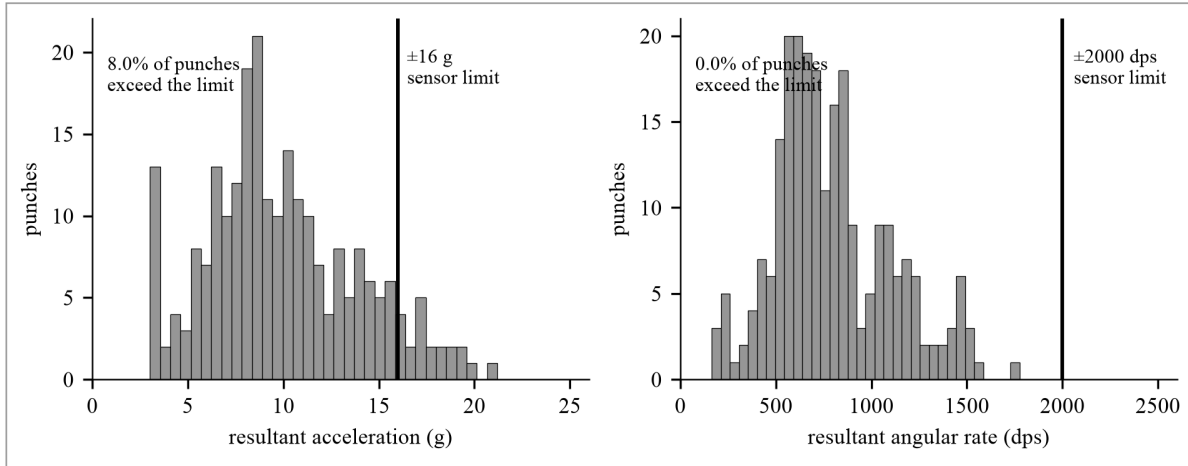


Figure 3. Peak resultant acceleration and angular rate per punch, against the development sensor’s configured ranges (LSM6DS3TR-C at ± 16 g, ± 2000 °/s). 238 punches, one athlete. Acceleration exceeds range on 8.0 per cent of punches and 2.9 per cent have a sample sitting on the rail; angular rate peaks at 1776 °/s and never reaches its limit. Both are **lower bounds**: at the delivered 33 Hz, a sub-100 ms event is sampled perhaps three times, so true peaks are almost certainly higher than those recorded.

Two signals do two jobs. **Kinematics** — trajectory, rotation, guard — drive classification and technique scores and need gyro range and rate. **Impact** — force and snap — drives power scores and needs a high-g accelerometer at a high sample rate. The development part, an LSM6DS3TR-C in a Seeed XIAO nRF52840 Sense, serves the first adequately and the second not at all.

Our own measurement sets both specifications. Eight per cent of punches exceed ± 16 g outright, so the high-g channel is justified directly by captured data rather than by a figure quoted from elsewhere. Angular rate peaked at 1776 °/s across 238 punches, inside the ± 2000 °/s range, so the ± 4000 °/s part is specified for headroom rather than necessity — a deliberate margin, since both numbers are measured at 33 Hz and the new board’s full-rate capture will resolve the true peaks. The wider gyro costs nothing at this BOM position and removes the question entirely.

4.2 Board architecture

A fully custom, company-specific board — not a XIAO carrier. The GPIO map deliberately mirrors the XIAO nRF52840 so existing firmware ports with only a new board definition.

Table 1. Principal components and the reasoning. Prices are October 2026 single-unit estimates and move with stock.

Function	Part and rationale	Est. \$
BLE MCU	Pre-certified nRF52840 module (Ebyte E73-2G4M08S1C, 18×13 mm, FCC ID 2ALPH-E73). A module avoids RF design and full intentional-radiator certification; the product carries “Contains FCC ID”. Alternative: Raytac MDBT50Q.	5.50–7.55
6-axis IMU	LSM6DSR (± 16 g, ± 4000 °/s) preferred over LSM6DSO32 (± 32 g, ± 2000 °/s); same LGA-14 2.5×3 mm footprint, so the choice can be deferred to BOM time (<i>verify pin compatibility</i>).	3.25–4.55
High-g accel	ADXL372 (± 200 g, 3200 Hz, anti-alias filter, 512-sample FIFO, 1.6 V min). Chosen over H3LIS331DL: ± 400 g but 1 kHz maximum and no FIFO.	12.00–18.50
Bus	SPI at 8 MHz. I ² C at 400 kHz cannot sustain both sensors at full rate. FIFO watermark interrupts with DMA burst reads.	–
Power	≈ 150 mAh protected LiPo into nRF52840 VDDH; internal REG0 at 3.3 V powers the sensors (≈ 1 mA total, well inside REG0’s budget), so no external regulator. MCP73831 charger at ≈ 100 mA.	≈ 1.00
Storage	Optional W25Q128 for event-windowed logging. Raw streaming at ≈ 40 kB/s fills 16 MB in ≈ 400 s, so logging must be windowed.	optional

A future single-chip option, the LSM6DSV320X (± 320 g high-g channel with a ± 4000 °/s gyro), could replace both sensors and remove most of the sensor cost (*verify price and stock*).

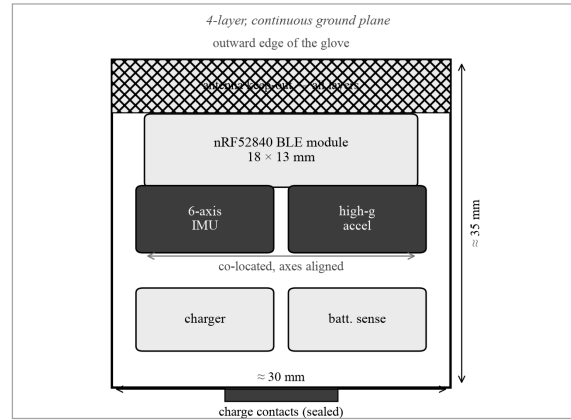


Figure 4. Board floorplan. The module antenna sits at the board edge with a copper keep-out on all layers and faces outward on the glove. The two sensors are co-located at the rigid centre with axes aligned, which is what makes clean fusion between them possible. Power and charge contacts occupy the bottom edge. Four layers: a continuous ground plane eases LGA escape, and at this size a 4-layer board costs about the same as 2-layer.

Environmental and safety. Conformal coating is not waterproofing. Development uses a silicone conformal coat with masked contacts inside a sealed TPU shell with a gasket; production moves to a moulded enclosure or overmould. Potting is deferred until the design is frozen — it prevents rework, stresses the MEMS parts and detunes the antenna. Charging uses pogo or magnetic contacts with a reverse-polarity Schottky and TVS; battery voltage is never exposed on a pad, because sweat electrolysis corrodes it. The cell mounts at the wrist cuff or back of the hand and never on the strike surface, inside a rigid shell, and only UN38.3-certified cells are used. Body-worn use requires an RF exposure (SAR exclusion) assessment before sale.

Toolchain. KiCad 10 with schematic-as-code via SKiDL^[10], an AI-assisted layout plugin^[9] for netlist import, ERC/DRC and exports, and kicad-cli for checks and Gerber/drill/position output. Footprints are pulled with easyeda2kicad against LCSC part numbers and verified against datasheet land patterns. Autorouting is used only for low-speed nets; power, USB and antenna-adjacent routing are done by hand. The division of labour is deliberate: automation handles the netlist, footprints, DRC triage and exports; a human owns placement and RF/power review.

5 Scoring: rubric as the interpretable layer

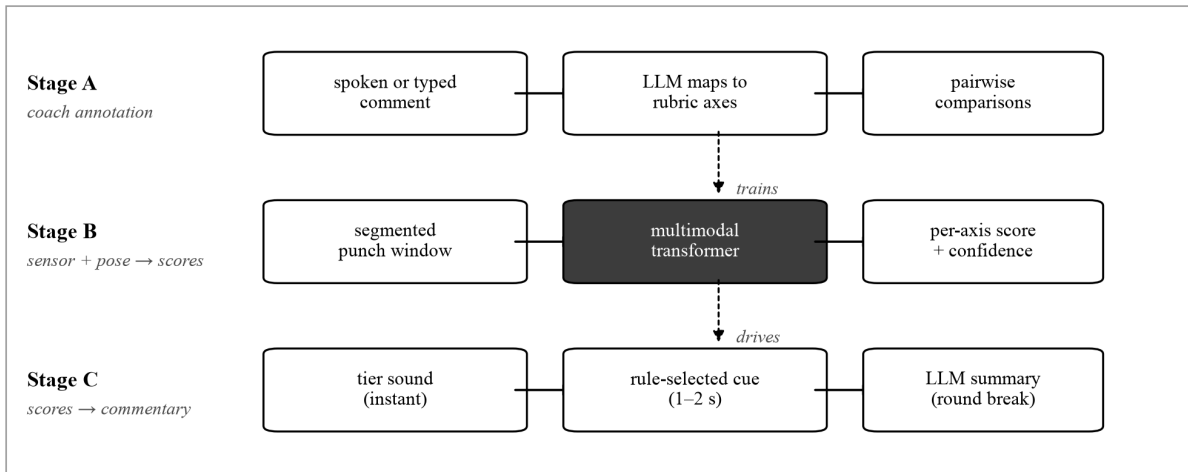


Figure 5. The three stages. Stage A turns coach judgement into rubric scores; Stage B learns to predict those scores from sensors and pose; Stage C turns predicted scores into speech at three different latencies. The transformer is the only learned component in the live path.

Action quality assessment is moving from single scores toward descriptive coaching. TechCoach^[1] introduced descriptive action coaching — a score together with what was done well and what to improve — and found point-level coach commentary a stronger training signal than scores alone, using an alignment loss between visual features and commentary embeddings. MTL-AQA showed that jointly learning recognition, scoring and commentary generalises better than any single task. In boxing specifically, BoxingPro^[3] drove an off-the-shelf language model from structured sensor-derived prompts and was rated above 4 out of 5 by professional boxers.

Stage A — coach annotation. The rubric is per punch type, four or five axes for the MVP, scored 1–5: guard position of the non-punching hand, hip and shoulder rotation, full extension, retraction to the chin, and balance after the punch. Power and snap join once high-g data exists. Multi-axis labels are used rather than one overall score, because a single number cannot be turned back into a cue. The coach speaks or types natural comments and a language model maps them to axes, with the raw text retained as an auxiliary signal and the mapping validated against direct coach scores. Pairwise comparisons — which of these two hooks was better — are mixed in, because they are more consistent than absolute ratings, and trained with a ranking loss.

Stage B — sensors and pose to scores. Input is one segmented punch window; inertial channels and RTMPose keypoints enter as time-aligned modality tokens. Outputs are punch type, one head per rubric axis, and a confidence. Losses are per-axis ordinal regression plus ranking, with a TechCoach-style alignment to comment embeddings. Self-supervised masked reconstruction pretrains on unlabelled punches. Target size is 1–5M parameters, running in real time on the phone via CoreML. Confidence gating means no feedback is given when the model is unsure.

A note from our own results bears on sizing. On this corpus a 60-parameter feature classifier reached 0.693 while 1D CNN, InceptionTime^[7] and DeepConvLSTM^[6] variants of 190k–400k parameters reached 0.35–0.46, each hitting 1.000 training accuracy within three epochs. Capacity must follow the labelled corpus, not precede it.

6 The coaching agent

Table 2. Feedback at three latencies. Only the slowest involves a language model; the live loop is deterministic, because latency, verbosity and hallucination are all unacceptable between punches.

When	Latency	What	Generated by
Per punch	instant	Tier — perfect / good / okay — with a sound	threshold on axis scores
In round	1–2 s	One short spoken cue from a curated library of 30–40	deterministic rules
Round break	≈30 s	Two-sentence summary of the round’s score log and trend; sets the next round’s focus axis	LLM with tools
Session end	–	Short saved report against previous sessions	LLM

Feedback design follows motor-learning evidence rather than intuition, and the main consequence is counter-intuitive: **the system deliberately stays quiet**. Constant feedback creates dependence and harms retention, so cues fire on bandwidth — only when an axis drops below threshold — at most one every 15–20 seconds, addressing one fault at a time, and fading in frequency as averages improve. Cues use an external focus (“snap it back to the chin”) rather than an internal one, and streaks are reinforced.

At round breaks the agent is given tools rather than prose: it can request round statistics, compare against the previous round, and set the focus axis for the next one. The MVP covers three punch types — jab, cross, hook — four or five axes, and bag work only.

7 Data and annotation plan

Collection runs on the current development hardware, which already carries the kinematic axes the MVP rubric needs. The power and snap axis joins the rubric once the custom board delivers high-g capture, or earlier if a high-g breakout is added during collection.

Table 3. Annotation budget over three weeks. Volume estimates assume 150–250 punches per round and 1,000–1,500 per six-round session, across 10–15 boxers.

Task	Per item	Target	Coach hours
Boundary correction (auto-proposed)	3–5 s	5,000 punches	5–7
Rubric scoring	10–20 s	2,000–3,000	10–15
Pairwise comparisons	5–8 s	1,500 pairs	≈3
Spoken comments	20–30 s	300–500	3–4
Re-scored repeats (consistency)	10–20 s	≈150	≈1
Total			≈20–25

That is roughly 1.5 hours a day, five days a week, with sessions held under ninety minutes to limit fatigue. Four principles govern the data.

Humans mark **temporal boundaries only**; RTMPose generates every keypoint, which avoids a train/inference mismatch. Scoring accuracy is capped by segmentation precision, so clean start, impact and end timestamps matter more than label volume. Diversity beats volume — more boxers across more skill levels, with two or three held out entirely as a test set. Inertial and video streams are synchronised with a triple glove-clap at session start.

The volume principle is not a preference but a measurement. On our wrist corpus, accuracy saturates at roughly 85 clips per stance and is flat thereafter: additional punches from the **same** athlete add nothing measurable, while each new athlete opens ground none of the existing data covers. Annotation effort therefore follows people rather than repetitions — which is also the cheaper of the two.

Training cost is not the constraint. Mapping comments to scores takes minutes and a few dollars of API usage; pretraining on 20–40k windows is 1–3 hours on one consumer GPU; fine-tuning is 10–

30 minutes, and hold-one-boxer-out evaluation a few hours. The success metric is model–coach agreement **relative to the coach’s own consistency**, measured by the re-scored repeats, with a shipping bar of roughly 80 per cent of human consistency. Guard and retraction are expected to work first; rotation is hardest.

8 Manufacturing cost

≈\$24.50 per glove · ≈\$49 per pair at ≈1,000 units falling to ≈\$17 per glove (≈\$34 per pair) with a single combined sensor (<i>verify</i>) delivered with duties, two-chip: \$45–54 per pair

All figures are October 2026 estimates of **cost to manufacture**. They exclude salaries, tooling amortisation and freight variability, and they say nothing about what the product would sell for.

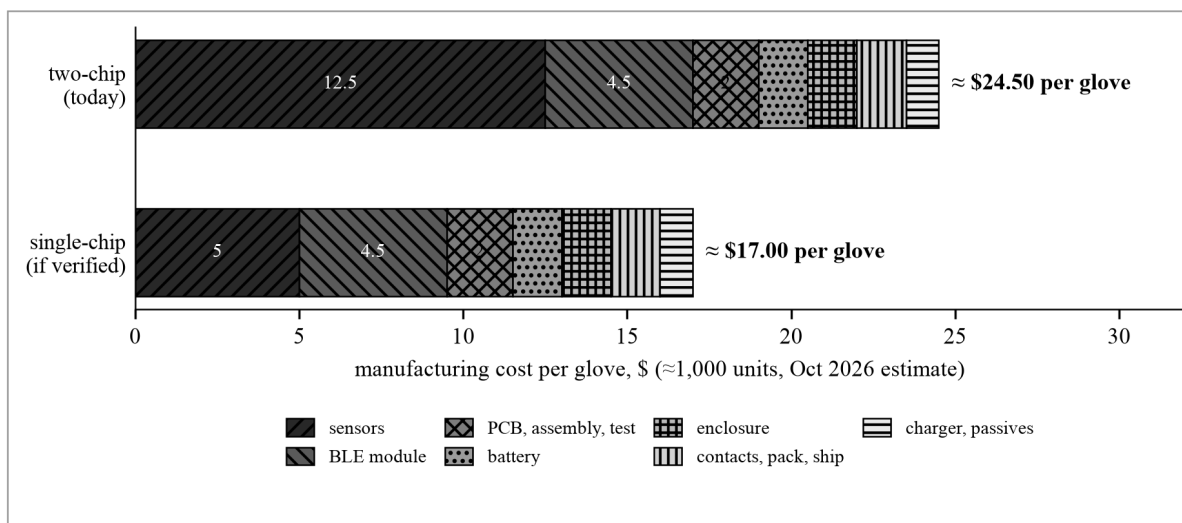


Figure 6. Per-glove manufacturing cost at approximately 1,000 units. One line dominates and only in the two-chip configuration: the sensor pair is 51 per cent of the bill of materials. That is why the single combined sensor is the highest-leverage open item in the hardware roadmap — it removes roughly \$7.50 per glove, about \$15 per pair, for no change to anything else on the board.

Table 4. Prototype, Rev A. Five assembled boards through a Chinese assembler, shipped to the United States.

Line	Est. \$
Parts per board, excluding battery	24–33
Components, five boards	120–165
PCB	2–7
Assembly setup and stencil	≈10
Extended-part fees	≈21
Shipping	20–30
Import duties	25–60
Five assembled boards, delivered	200–290
Reduced build: hand-place the high-g part, omit flash, assemble 2–3 of 5	170–220

Two cost drivers are structural rather than incidental. The high-g accelerometer is listed as standard-assembly-only, adding \$35–45 to a run that includes it, which is why hand-placing it is the single largest prototype saving. And de minimis no longer applies to China-origin shipments: effective rates on assembled boards ran about 12.5–35 per cent through 2026 and changed several times, so duties-prepaid checkout is used to avoid surprise brokerage. Section 301 exclusions lapse on 10 November 2026 unless extended (*verify current rates before ordering*).

Table 5. Production, per glove, at approximately 1,000 units. Roughly 15–20 per cent lower at 10,000.

Line	Two-chip	Single-chip
BLE module	4–5	4–5
Sensors	≈12.50	4–6 (<i>verify</i>)
Charger and passives	≈1	≈1
PCB, assembly, test	1.50–2.50	1.50–2.50
Battery	1–2	1–2
Enclosure	1–2	1–2
Contacts, packaging, shipping	1–2	1–2
Per glove	21–27	14–20
Per pair, before duties	42–54	28–40
Per pair, delivered with duties	45–54	–

One-time costs are enclosure tooling at \$3–15k, modular FCC plus body-worn RF exposure assessment at \$3–10k, a test fixture at \$0.5–2k, and a pilot run of 100–300 units at \$3–8k. Moving from the pre-certified module to a bare nRF52840 saves roughly \$1.50–2.50 per glove but adds \$10–20k of certification and real RF risk; it is only worthwhile at tens of thousands of units. Factory-installed batteries ship under easier rules than loose cells.

At prototype scale the custom board saves only \$1.50–3.50 per unit against the development module, which does not justify it on cost. It is justified because the board **is** the product: size, sealing, sensor placement and the ability to put a high-g channel next to the IMU are not available any other way.

9 Roadmap

Camera–IMU fusion is the core technical direction, and §3 is the argument for it. The camera supplies whole-body context the gloves cannot see — feet, hips, trunk rotation, guard relative to the head — which is precisely the information that separates hook from uppercut. The gloves supply what the camera cannot: high-rate wrist dynamics, impact, and continuity through occlusion and motion blur. Fusion happens inside the scoring transformer as modality tokens, with inertial-only and camera-only fallback modes for gyms without a camera setup. Time sync uses hardware time-stamps from the sensor FIFO, Bluetooth clock alignment, a visual sync event and drift correction; pose additionally calibrates glove orientation to the body frame at session start, which would have prevented the frame mismatch described in §3.

Hardware moves to the single combined sensor if it verifies, then a sealed pogo-charging Rev B, a fuel gauge, a moulded enclosure, and over-the-air updates; a bare-chip design only at high volume.

Models add the power and snap axis once high-g data exists, extend to all six punch types and combinations, treat shadowboxing and mitt work as separate modes because impact changes the signal, bring in a second coach to measure and reduce single-coach bias, and personalise per athlete.

Product adds a gym dashboard for classes of fifteen to thirty, agent-selected drills, branded coach voices and progress tracking.

10 Open questions and how they are addressed

Each item below is paired with the work that closes it. Nothing here is blocking; the sequence in §9 is ordered so the highest-leverage items come first.

- **Multi-athlete validation** is the next milestone. Results to date come from one athlete, and the learning curve in §7 shows why that is the right place to spend effort next: additional punches from the same person add nothing, while each new athlete opens ground none of the existing data covers. The annotation plan is already built around holding two to three boxers out entirely.

- **Coach consistency** is measured directly, by re-scoring repeated clips, and the shipping bar is set relative to that figure rather than to an absolute score. A second coach joins in the model roadmap to quantify and reduce single-coach style.
- **Segmentation precision** sets the ceiling on scoring accuracy, which is why boundary correction gets its own line in the annotation budget and is auto-proposed rather than hand-marked from scratch.
- **Full-rate capture** arrives with the custom board. The SPI bus, FIFO watermark interrupts and DMA burst reads in §4.2 exist to deliver the full sample rate, which will both enable the power axis and resolve true signal peaks.
- **Battery safety and RF compliance** are designed for from the start: cells are UN38.3-certified and mounted away from the strike surface, and the pre-certified module keeps certification to a modular filing plus a body-worn RF exposure assessment.
- **Landed cost** moves with tariff policy, which changed several times during 2026; duties-prepaid checkout fixes the cost at order time, and the single combined sensor reduces exposure by cutting the dutiable bill of materials.
- **To confirm before release:** single-chip sensor pricing and stock; LSM6DSR and LSM6DSO32 pin compatibility; assembler part classifications; current duty rates.
- **Intellectual property.** Publication is a disclosure. The United States allows a one-year grace period; most other jurisdictions require absolute novelty, so any patentable hardware or method should be filed before this document is released publicly, with counsel reviewing it first.

11 References

- [1] TechCoach — descriptive action coaching. arXiv:2411.17130
- [2] Action quality assessment survey. *Journal of Big Data*, 2026.
- [3] BoxingPro. *Electronics* 14(21):4155, 2025.
- [4] SPAR — 8-IMU garment, three-tier explanations. arXiv:2609.30753
- [5] CoachMe. arXiv:2509.11698
- [6] Ordóñez and Roggen, DeepConvLSTM. *Sensors*, 2016.
- [7] Ismail Fawaz et al., InceptionTime. 2020.
- [8] Jiang et al., RTMPose. 2023. Pose backbone, used as released.
- [9] mcp-server-kicad — AI-assisted KiCad layout.
- [10] SKiDL — schematic as code.
- [11] Component datasheets: ADXL372 (Analog Devices), LSM6DSR / LSM6DSO32 / LSM6DSV320X (ST), nRF52840 product specification (Nordic).

Attribution. BoxingVI, RTMPose and ST-GCN++ are third-party work used under their respective licences and credited to their authors; BoxingVI’s licence should be reviewed before any commercial deployment of models trained on it. Our contribution is the capture rig and protocol, the wrist corpus, the conversion and training pipeline, the detection and calibration design, the board specification, the on-device integration, and every measurement reported here.

Reproducibility. Every figure is regenerated from data on disk by `scripts/whitepaper_figures.py` and `scripts/whitepaper_diagrams.py`. No value is transcribed by hand except the camera confusion matrix, quoted from `MODELS.md` and marked as such in the source.

Estimates. All costs and timelines are October 2026 estimates. Items marked (*verify*) are unconfirmed at the time of writing.