INTERNSHIP FINAL REPORT

Deploying and Hardening a Thermal Fall-Detection System on Edge Hardware

Real-time inference on Jetson Orin · 45-session thermal dataset · six detector versions benchmarked at IoU@0.5

Phuc Nguyen Hong (VinUniversity)
Base system: TaFall - Chengxiao Li, Xie Zhang, Chenshu Wu (The University of Hong Kong)
August 2026

Falls Are Common and Harmful - but Hard to Monitor Privately

28–35% of adults 65+ fall at least once per year

20–30% of falls cause mild to severe injury

10–15% of all emergency department visits

15–55% restrict daily activity from fear of falling

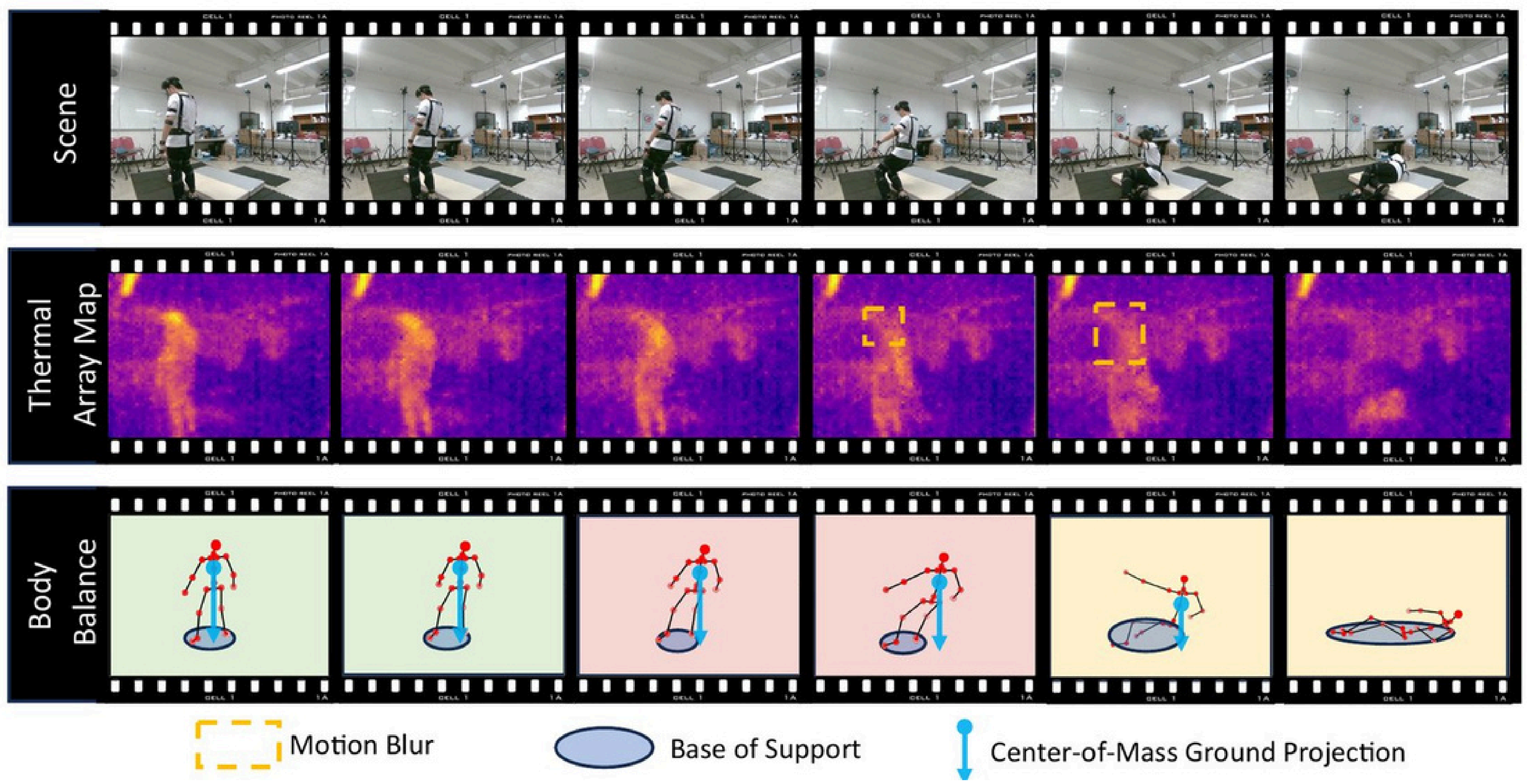
Most falls happen in bedrooms and bathrooms - exactly where cameras are unacceptable.

Sensing modality	Privacy	Accuracy	Key limitation
Wearable (IMU)	High	High	Adherence: discomfort, charging, forgetting to wear
RGB camera	**Low**	High	Unacceptable in bedrooms / bathrooms; fails in the dark
RF / radar	Medium	High	Coarse spatial resolution → high false-alarm rate
Thermal array (prior work)	High	**Low**	No pose or balance modelling - silhouette heuristics only
Thermal array + pose (this work)	**High**	**High**	62×80 px, works in total darkness, no identity captured

Deployment bottleneck: even a 2% false-alarm rate over 1-minute windows produces dozens of false alerts per day - caregivers stop trusting the system.

STARTING POINT

The Baseline: TaFall, a Balance-Informed Thermal Detector



A fall in three views: RGB scene, 62×80 thermal map, and the balance state the system actually reasons over. Green = stable balance, red = loss of balance, amber = ground impact. Source: Li et al., TaFall, arXiv:2604.09693, Fig. 1.

Core idea

- A fall is a transient loss of balance (WHO definition), not simply a height drop - so the model tracks the centre of mass against the base of support.
- Sensor: Meridian MI48 / MI0802M6S thermal array + microcontroller, about USD 18 in parts.

Processing pipeline

1. **IRA sequence**
   62×80 thermal, sliding window
2. **Person detector**
   IRATrackerDetector, CenterNet-style 3D-CNN
3. **Pose estimator**
   17-joint 2.5D skeleton from the masked crop
4. **Balance features**
   CoM vs. base of support; motion from thermal blur
5. **Fall classifier**
   ST-GCN + Transformer → Normal / Fall

Meridian MI48

Published results - lab's work, not my contribution

Detection rate, 3,000+ falls

False-alarm rate, benchmark

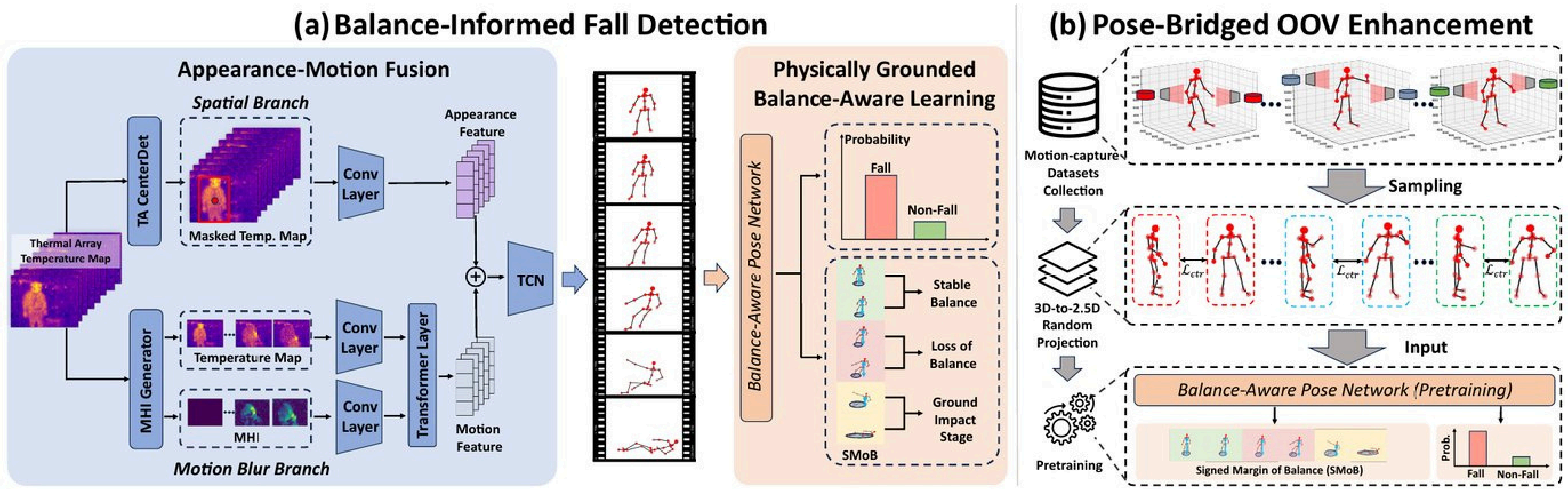
Field trial, 2 care facilities

27 days

Long-term deployment FAR

0.00126%

TaFall Architecture, as Published



Source: Li, Zhang, Zhu, Jiang & Wu, “TaFall: Balance-Informed Fall Detection via Passive Thermal Sensing”, arXiv:2604.09693, Fig. 3.

- (a) Balance-Informed Fall Detection - an appearance–motion fusion module recovers a 2.5D pose sequence from the thermal maps, which a balance-aware pose network turns into a balance state and a fall decision.
- (b) Pose-Bridged OOV Enhancement - the pose network is pretrained on public motion-capture data projected into many 2.5D viewpoints, which is what suppresses false alarms on unseen everyday behaviours.
- My internship touched only the person-detector stage inside (a) and the runtime that executes this graph - the balance modelling itself is unchanged.

SCOPE

Three Workstreams Over the Internship

1

Real-time edge deployment

Re-architected the inference pipeline so the full detect → pose → classify chain runs live on an NVIDIA Jetson Orin with a browser video feed.

12 → 20.8 FPS

display latency 3 s → < 0.3 s

2

New thermal dataset

Collected and auto-labelled a synchronised thermal + RGB + depth dataset, extended in a second phase to cover two- and three-person scenes.

45 sessions · 91,800 frames

16 of them multi-person

3

Detector fine-tuning

Trained and benchmarked six detector versions at IoU@0.5, isolating what each change actually bought.

40.7% → 1.5% FPR

false positives on empty frames

All three share one goal: make the system trustworthy enough to leave running unattended.

Rebuilding the Inference Loop for Real-Time Operation

Sensor thread
MI48 @ 25 FPS,
auto-discovered COM port,
8-attempt reconnect

›

Sliding window
T = 20 frames
(was 40)

›

Streaming detect + pose
d2pose_streaming +
SkeletonStreamer cache

›

Fall classifier
AMP FP16,
inference every 4th frame

›

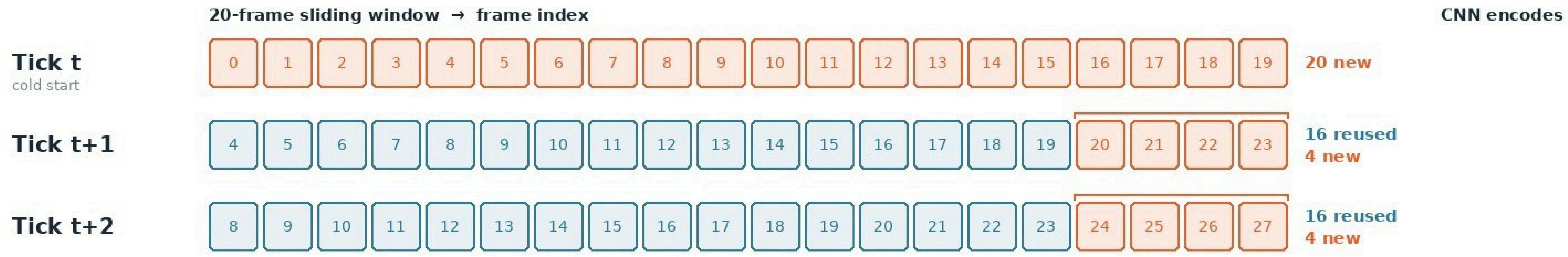
Track + serve
IoU tracker with EMA
→ Flask MJPEG stream

Four changes account for almost all of the speed-up

Change	Before	After	Effect	Why it is safe
Window length	40 frames	20 frames	−50% buffer memory	20 frames still spans the full fall event at 25 FPS
Inference stride	every frame	every 4th frame	−75% forward passes	A fall lasts ~1 s; 6 inferences per second is ample
CNN feature cache	none	SkeletonStreamer	**−80% CNN encode calls**	**Bit-identical features - next slide shows why**
Numeric precision	FP32	AMP FP16 (CUDA)	≈1.5× faster on Orin	Loss computed in float32; inference-only change

Everything above is an engineering change - the trained weights are untouched, so accuracy is unaffected.

Why the Feature Cache Is Safe as Well as Fast

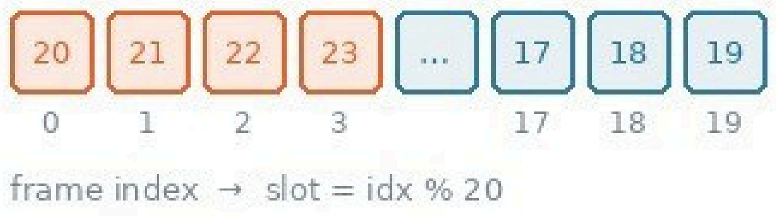


16 of every 20 encodes are cache hits

→ 80% fewer CNN forward passes per inference tick

Why the key can never collide

The cache is keyed by slot = abs_frame_idx % 20 . A window only ever holds 20 consecutive indices, so no two frames inside one window can map to the same slot — the oldest frame's slot is overwritten by the newest at exactly the moment it leaves the window.



Measured effect on the Jetson Orin: CNN encode calls per inference tick fall from 20 to 4. Combined with the stride change this is the single largest contributor to the 12 → 20.8 FPS result; the motion-history (MHI) feature is likewise computed once and reused for the lifetime of the stream.

From a Laggy Demo to a Usable Live Feed

sustained on Jetson Orin (was ~12 FPS)

20.8 FPS

end-to-end display latency (was ~3 s)

< 0.3 s

Configuration	Throughput	Display delay
v1 - window 40, stride 1	~12 FPS	~3 s
v2 - window 20, stride 4	**20.8 FPS**	**< 0.5 s**
v2, viewed over SSH port-forward	~20 FPS	~2 s
v2, viewed over direct LAN IP	**20+ FPS**	**< 0.3 s**

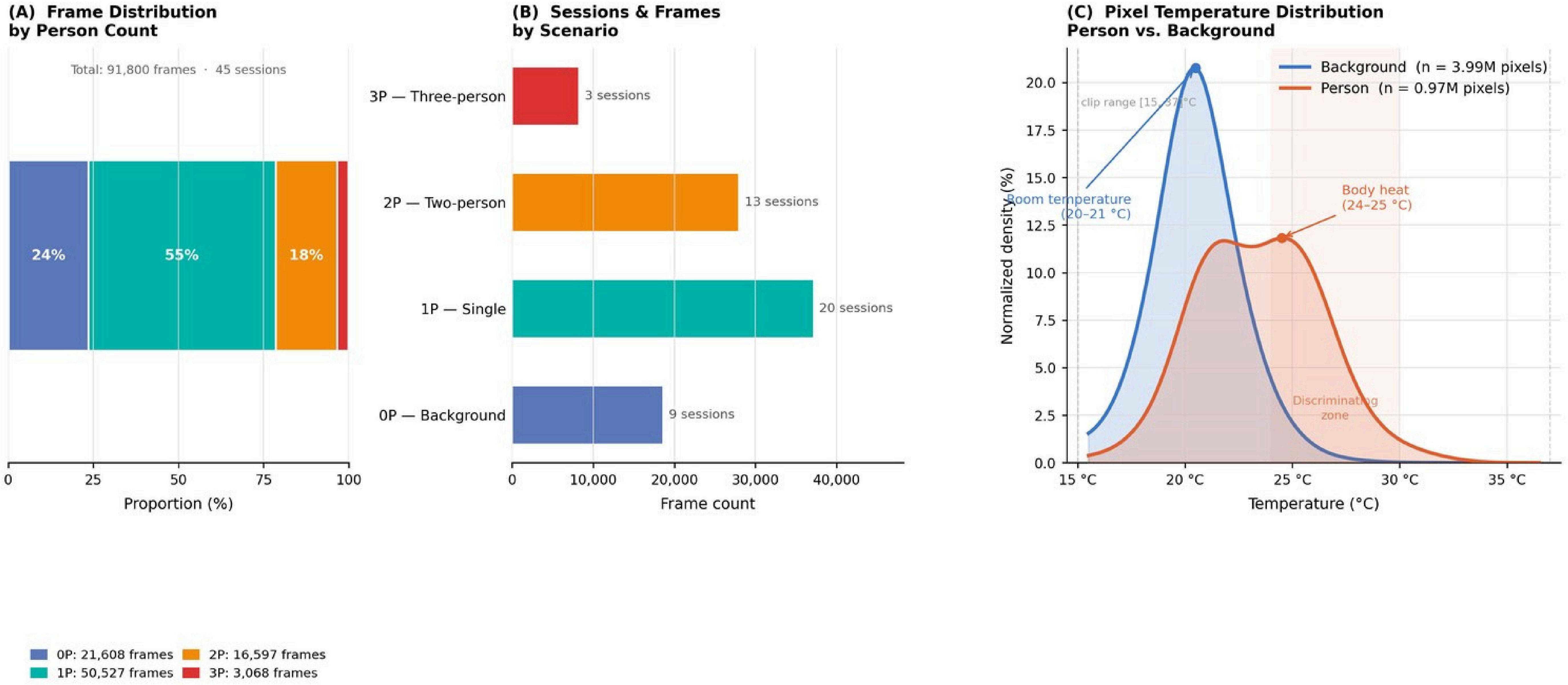
A Thermal Dataset Built for the Scenes That Actually Fail

45
recorded sessions, ≈51 minutes of capture

91,800
IRA frames - 62×80 px at 25 FPS

16
multi-person sessions (13 two-person, 3 three-person)

9
background-only sessions: 21,608 empty frames



Phase 1 → Phase 2: 35 → 45 sessions, and multi-person coverage went from a single session to 16. Session length ranges from 410 to 5,850 frames (median 1,900).

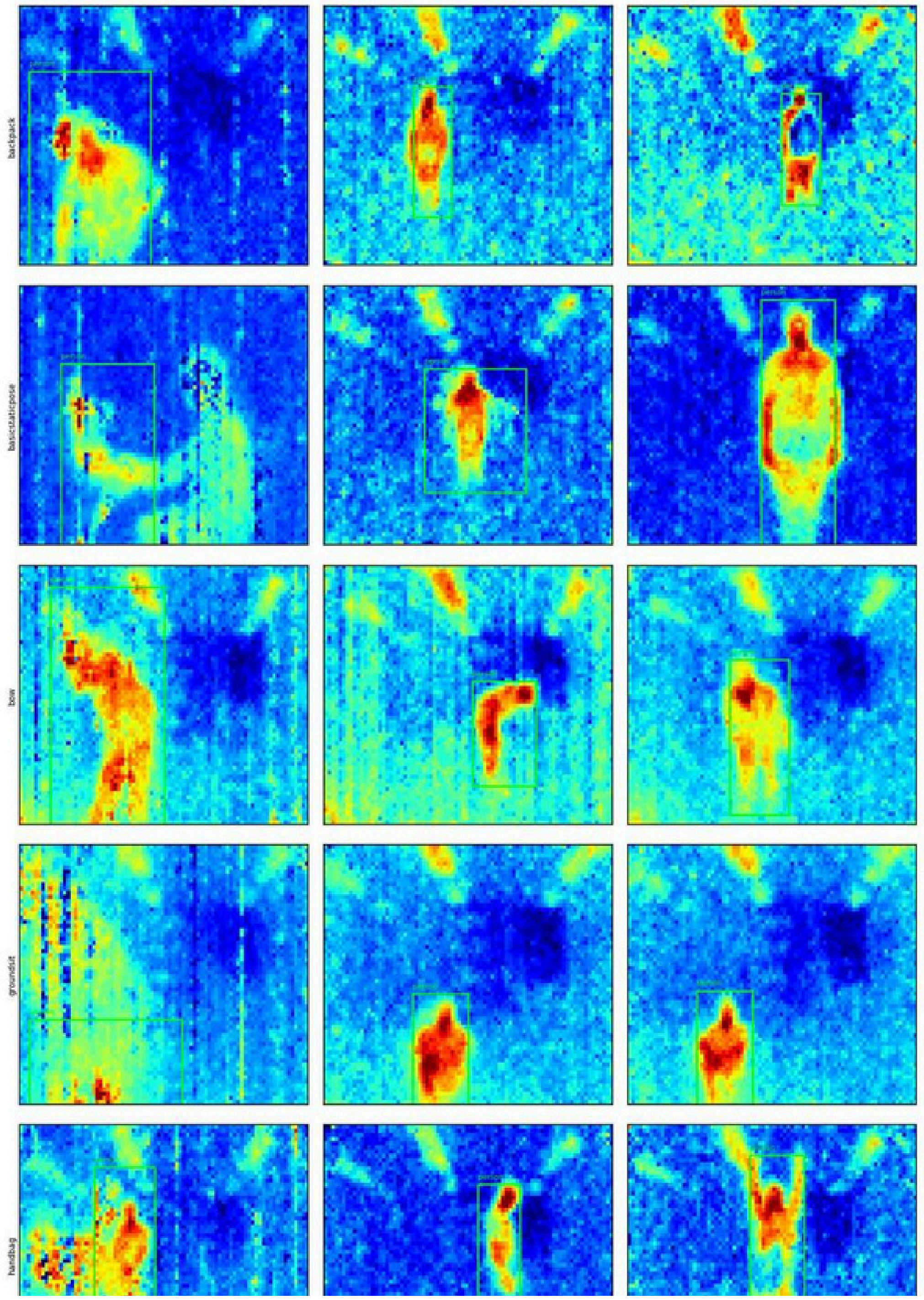
Labelling 91,800 Frames Without Annotating Any By Hand

1. **Detect in the easy modality**
   Run YOLOv8n-Pose at conf ≥ 0.55 on the RGB frame whose Unix timestamp is nearest to each thermal frame. RGB is high-resolution, so off-the-shelf detection is reliable there.

2. **Project RGB → thermal**
   Map each box from 480×640 RGB space into 62×80 IRA space using depth-calibrated field-of-view alignment, then apply the vertical flip that matches the runtime convention.

3. **Reject bad projections thermally**
   Discard any projected box whose maximum temperature is below 24 °C. A box that lands on a cold surface cannot contain a person - this catches residual misalignment automatically.

4. **Label empty scenes explicitly**
   The nine background sessions skip YOLO entirely and receive empty box lists, so the detector is trained on what "no person" looks like rather than never seeing it.



Auto-generated boxes overlaid on thermal frames, verified visually before training.

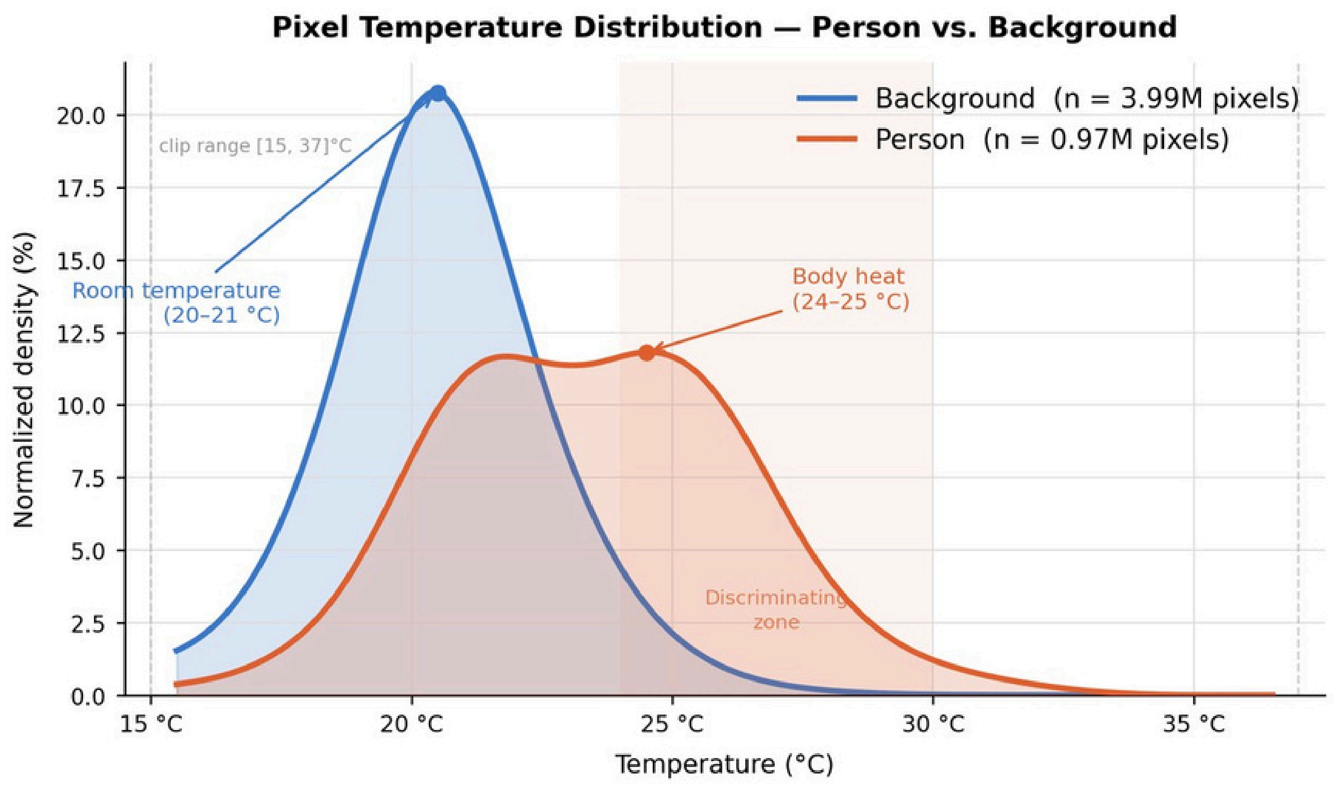
Manual annotation of 91,800 thermal frames was never realistic. Outputs were checked on rendered overlay videos, not assumed correct.

Six Versions, and Only One Change Really Mattered

The starting point

Version 1 fired a false positive on 40.7% of frames that contained no person at all. Three rounds of extra data and retraining only brought that to 34.3%.

The actual culprit was preprocessing, not the model



Clipping the input to [20, 30] °C truncated both tails of the person distribution - a body at 32 °C and a cool limb at 18 °C were both flattened into the background. Widening to [15, 37] °C fixed it.

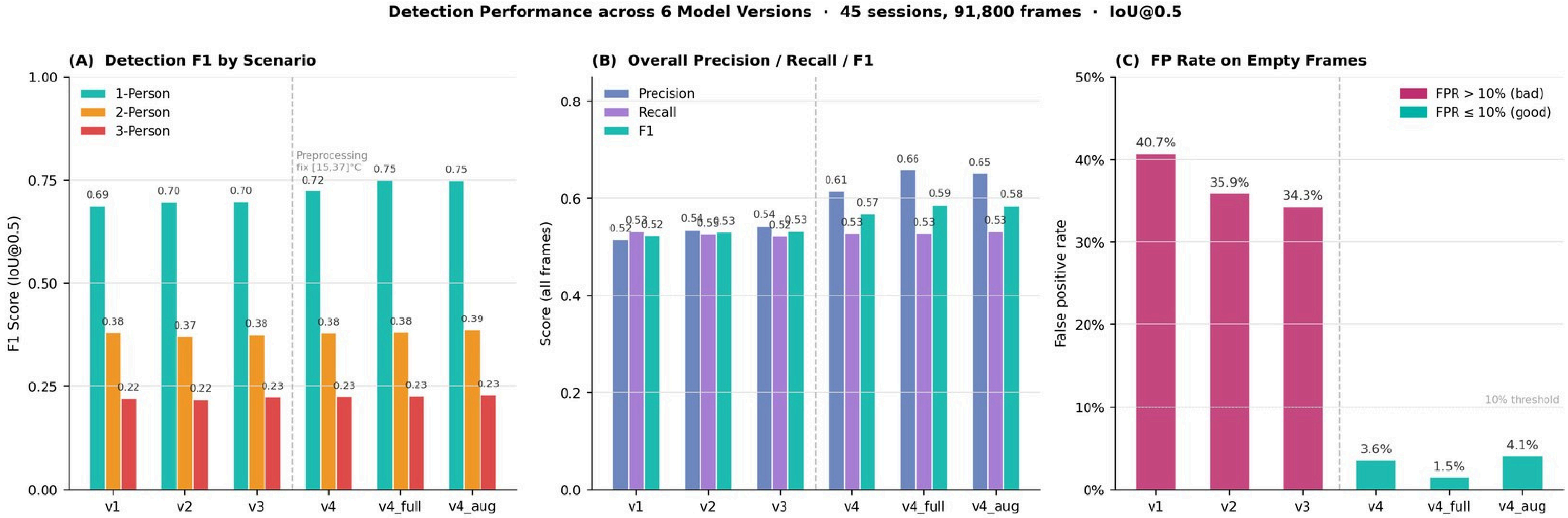
What each version changed

v1 → v3	**More data, more finetune iterations** 1P F1 +0.010, FPR 40.7% → 34.3%. Slow, unconvincing progress.
v3 → v4	**Preprocessing clip [20,30] → [15,37] °C** FPR 34.3% → 3.6% - a 10× drop from one line of preprocessing.
v4 → v4_full	**Unfreeze the whole network** 1P F1 +0.026, FPR 3.6% → 1.5%. Now safe: 45 sessions, not 34.
v4_full → v4_aug	**Six augmentations added** F1 unchanged, robustness to sensor and viewpoint shift improved.

Production model: v4_aug, validation loss 0.3970, detection threshold 0.36.

AdamW · cosine schedule · batch 4 · 5-frame sequences · AMP float16 · focal heatmap loss + 0.1 × L1 size, computed in float32 to stop log(0) → NaN

Measured at IoU@0.5 Across All Six Versions



40.7% → 1.5%

False positives on empty frames. The production model (v4_aug) sits at 4.1%, trading a little for robustness.

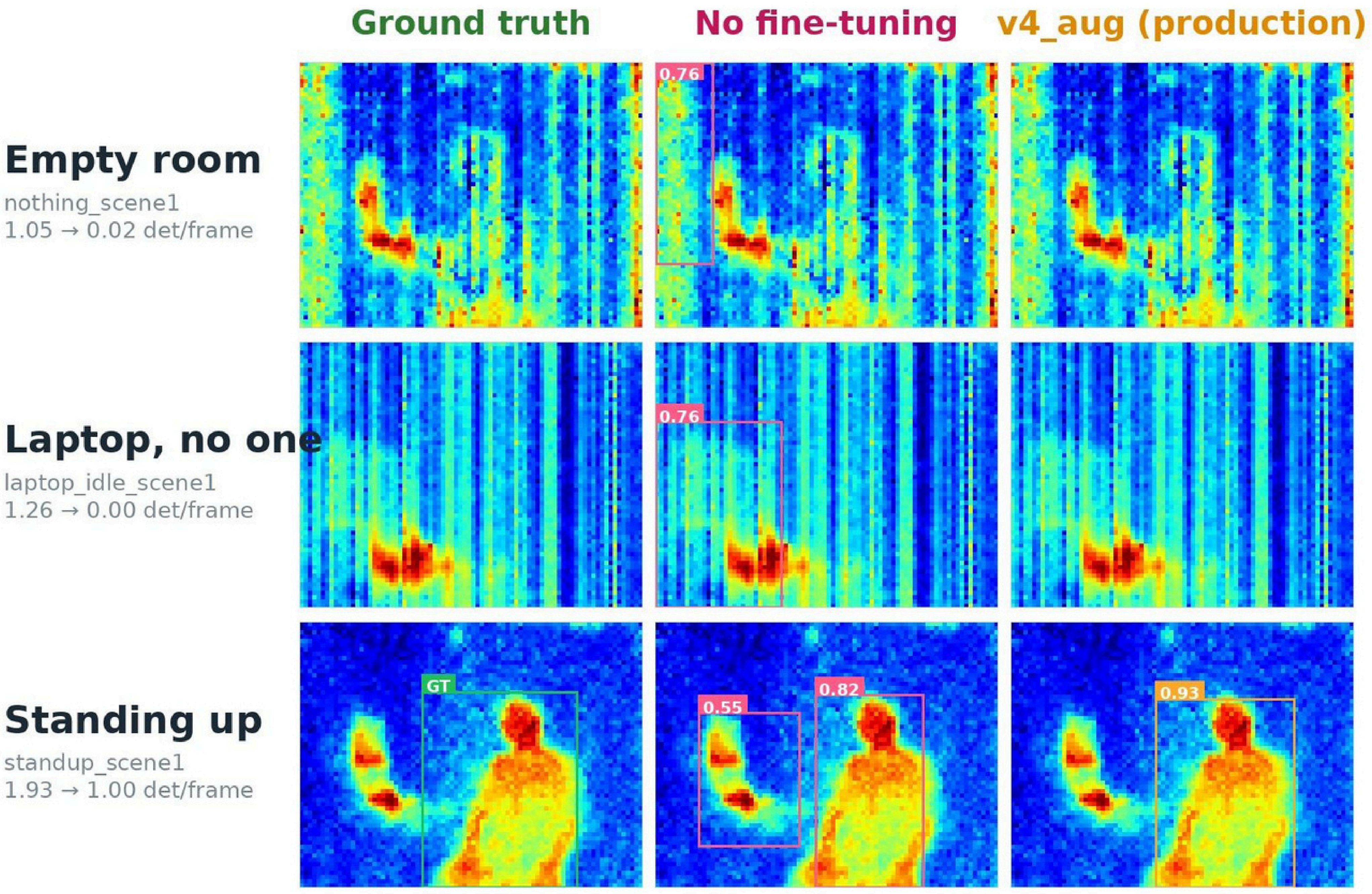
0.688 → 0.749

Single-person F1. Detection quality rose while false alarms fell - the model became selective, not blind.

In-distribution

Split is frame-level 85/15, so every session appears in training. This is not a held-out test set - see limitations.

No Fine-Tuning vs. the Production Model



Row by row

Empty room

The released weights fire at 0.76 on an empty room. v4_aug goes silent: 1.05 → 0.02 detections per frame.

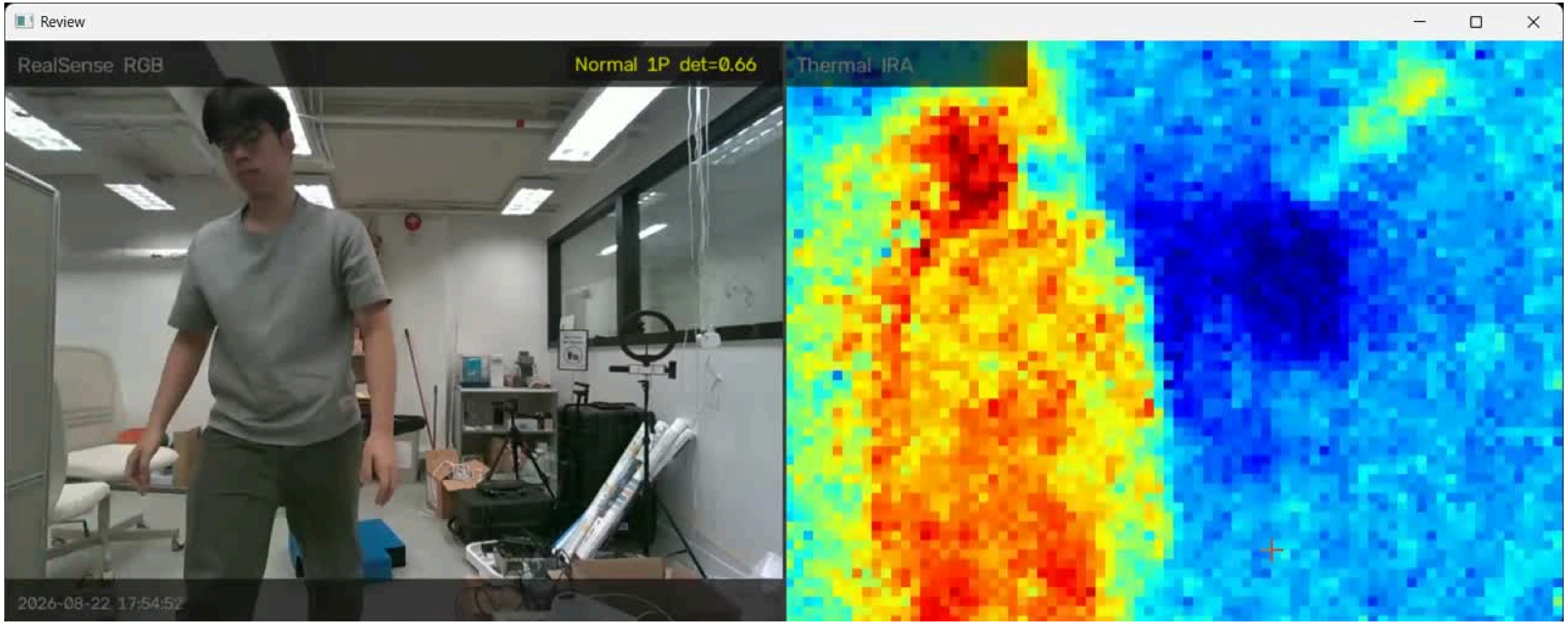
Laptop, no one present

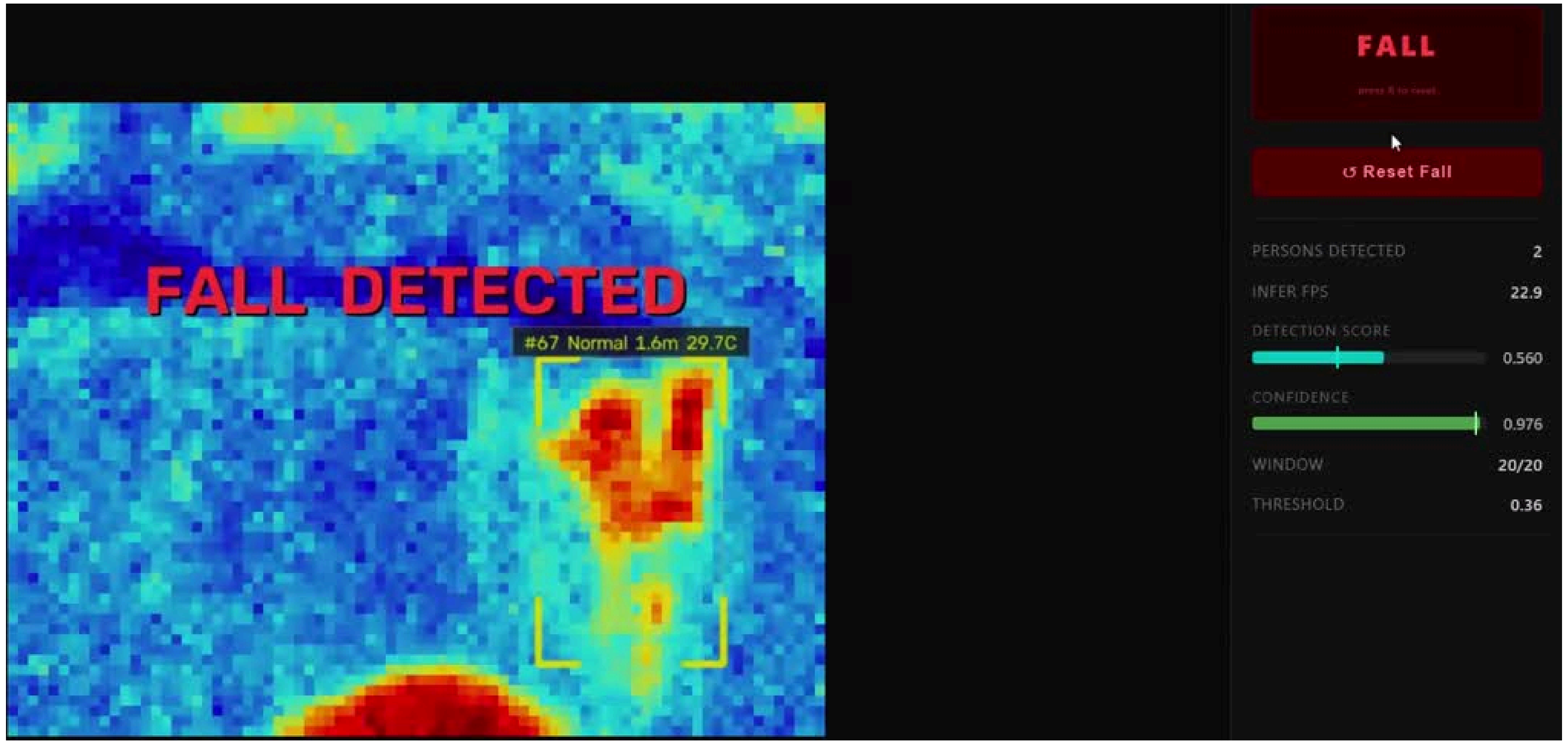
A laptop's heat plume is boxed at 0.76 - the exact failure the geometric filter was written for. The tuned model does not need that filter here.

Standing up

Two boxes become one: the phantom on the warm object at left disappears and the person's score rises from 0.82 to 0.93.

DEMO


Review
RealSense RGB
Normal 1P det=0.66
Thermal IRA
2026-08-22 17:54:52


FALL DETECTED
#67 Normal 1.6m 29.7C
FALL
↺ Reset Fall
PERSONS DETECTED 2
INFER FPS 22.9
DETECTION SCORE 0.560
CONFIDENCE 0.976
WINDOW 20/20
THRESHOLD 0.36

CONCLUSION

Delivered

20.8 FPS	live thermal detection on Jetson Orin, latency under 0.3 s
45 sessions	91,800 auto-labelled frames, 16 of them multi-person
40.7% → 1.5%	false positives on empty frames, across six benchmarked versions
+1.07 °C	measured and corrected offset between the two sensor units

Thank you - questions welcome.

Acknowledgements: I would like to express my sincere gratitude to Prof. Wu, Xie Zhang, and Chengxiao Li for their invaluable support during my internship.

BACKUP SLIDES

Supporting detail for questions

Multi-Person Is Where the Architecture Runs Out

Two-person F1 at IoU@0.5 (v4_aug)

0.388

Three-person F1 at IoU@0.5 (v4_aug)

0.230

What I fixed

A second person produced a real but weak heatmap peak - 0.15 to 0.25, below the 0.36 detection threshold - so it was silently dropped.

Parameter	Before	After
NMS search window	±1 feature px	±2 feature px (5×5)
Secondary peak threshold	0.36	0.18 (thresh × 0.5)
Box merge condition	overlap or side-by-side	IoU > 0.30 only

Why it stops there - an architecture ceiling, not a training problem

The detector predicts a 16 × 20 heatmap. At that resolution one person already occupies several cells, so three people produce Gaussian peaks that overlap before non-maximum suppression ever runs.

More training data does not separate peaks that the output grid cannot represent. 2P F1 ≈ 0.39 and 3P F1 ≈ 0.23 are the ceiling of this head, and every version from v1 to v4_aug sits within 0.02 of it.

Fixing it means changing the head - a finer output grid, or an anchor-free multi-peak formulation - not collecting more three-person sessions.

The fix above also has a cost: a secondary threshold of 0.18 is permissive by design, and it is part of why v4_aug's empty-frame false-positive rate (4.1%) sits above v4_full's (1.5%).

Worth saying out loud: three-person scenes are 3.3% of the dataset, so this ceiling barely moves the headline F1 - which is exactly why reporting the per-scenario breakdown matters.

BACKUP

The Six Augmentations in v4_aug

Augmentation	Parameters	What it is meant to buy
Horizontal flip	p = 0.5, box coords updated	Invariance to which way the room faces
Random translate	±5 px in x and y, fill = median	Tolerance to sensor mounting position
Temperature shift	±1.5 °C uniform	Sensor-to-sensor variation - the +1.07 °C offset is inside this range
Contrast jitter	×[0.8, 1.2], p = 0.8	Different ambient room temperatures
Gaussian noise	σ = 0.3 °C, p = 0.8	Sensor read noise
Random erasing	p = 0.4, 10–20% area, consistent across T	Occlusion by furniture; the temporal consistency matters

Why keep v4_aug when v4_full has a lower false-positive rate

On the in-distribution validation set the two are within 0.001 F1 of each other (0.586 vs. 0.585), and v4_full's 1.5% empty-frame FPR beats v4_aug's 4.1%. The augmented model was chosen for deployment because the temperature-shift and noise augmentations directly model the domain gap between the collection sensor and the Jetson sensor - a gap the validation set cannot see. That is a judgement call, and it is the one claim in this report that is not yet backed by a measurement.

Domain Shift Between the Two Sensor Units

The unit used for data collection (COM14) and the unit deployed on the Jetson (COM7) do not read the same temperature for the same scene. Left uncorrected, every threshold in the pipeline inherits a systematic bias.

Measurement	Value	Interpretation
Mean offset, COM7 − COM14	+1.07 °C	COM7 reads consistently warmer
Spatial std. dev. of the difference	0.40 °C	Offset is uniform, not a lens or vignetting artefact
Pixels where COM7 > COM14	99.7%	Systematic, not noise
Maximum per-pixel difference	+3.66 °C	Worst case, localised
Correction applied	**set_offset_corr(−1.07)**	**Applied on COM7 at start-up**

Takeaway: a model trained on one physical sensor unit is not automatically valid on another unit of the same part number. Calibration is required whenever the collection and deployment sensors differ.

Non-Model Fixes for Non-Human Heat Sources

1

Laptop detected as a person

A laptop vent produces a small, roughly square blob at ~33 °C - thermally indistinguishable from a torso at low resolution.

Fix

Connected-component gate on geometry: reject blobs with area < 120 px and height/width < 1.3. A standing person is always tall and narrow; a vent is not.

2

Static warm objects after warm-up

Radiators and equipment reach body temperature and stay there, so a static thermal threshold cannot separate them.

Fix

Ten-second background model at start-up; pixels within 2 °C of the background are neutralised, while anything deviating more is preserved as a moving body.

3

Degree sign rendered as ??C

OpenCV's built-in fonts do not support Unicode U+00B0, so every temperature label was corrupted on the video overlay.

Fix

Substituted a plain ASCII C in the overlay text. Cosmetic, but it made the demo videos unreadable until it was fixed.

These sit alongside the fine-tuned detector, not inside it - cheap geometric and temporal constraints handle failure modes it is wasteful to spend model capacity on.

Anticipated Questions

Why thermal instead of RGB?	No identifiable faces, works in total darkness, and is unaffected by lighting changes - the only modality acceptable in a bedroom or bathroom.
Is 62×80 px enough?	A person spans 20–30 px of height, which is enough for silhouette and coarse pose. Low resolution is also the point: it is cheap, fast enough for edge hardware, and privacy-preserving by construction.
Why not just use YOLO?	YOLO is pretrained on high-resolution RGB; the domain gap to a 62×80 thermal array is too large. YOLO is used here only offline, on the synchronised RGB stream, to generate labels.
What hardware produced the 20 FPS figure?	Jetson Orin, single sensor, stride 4 and window 20. Running two sensors concurrently gives roughly 15 FPS.
Is sensor calibration mandatory?	Only when the collection and deployment units differ, which was the case here: +1.07 °C systematic offset would otherwise bias every temperature threshold.
Isn't a frame-level split leaking data?	In effect, yes. Neighbouring frames are highly correlated, so the 15% held out overlaps heavily with what was trained on. Every number in this report should be read as an in-distribution upper bound, and a session-level re-evaluation is the first item of future work.