

Identity Through Interruption: A Motion-Signature-First Decision Architecture for Multi-Object Re-Identification

Draft manuscript — prepared for submission to *Nature Machine Intelligence*

Sehaj Randhir Singh
Independent Research
sehajr.singhs@gmail.com

Abstract

Appearance-based person re-identification (Re-ID) fails exactly where identity matters most: visually identical targets, occlusion, and re-entry leave appearance embeddings uninformative or misleading. We present an identity architecture in which **motion is the primary signal**: an online, forward-only decision layer grows bounded uncertainty regions around missing targets, stages gate $\rightarrow$ verify $\rightarrow$ bind decisions that hold hypotheses rather than guess under ambiguity, fuses gait, trajectory style, and geometry, and tracks co-located identity groups through pile-ups without re-spawning them. On synthetic hard segments the system sustains **statistically zero identity switches (IDSW) across 20 seeds** — including an appearance-free ballistic target — where appearance-first baselines show up to $30\times$ more. On the MOT17 public-detection benchmark (all seven train sequences), the pose-free group-level mechanism transfers: identity switches fall 21% ($1,265 \rightarrow 995$) and hard-IDF1 rises from 0.300 to 0.324 — consistent across five of seven sequences (paired per-sequence $p = 0.156$ at $n = 7$) — while uncalibrated real appearance channels remain harmful to every tracker until calibrated. Under a real low-recall detector stream, a velocity-consistency re-acquisition gate closes the IDF1 gap to 0.7 pp on the most stress-affected sequence, and a detector-recall audit shows that no tracker — including appearance baselines — escapes the detection floor. All transfer constants can be self-calibrated from the detection stream alone (no ground truth), and a generalizable channel-calibration audit flags the appearance channels of two independent trackers as net-harmful at operating scale. The decision architecture — not the channel models — is the portable contribution.

Editorial summary

Person re-identification — deciding that the person seen now is the same person seen before — is usually done by comparing appearance: clothing, face, build. That works when people look different, and fails exactly when identity matters most: identical uniforms, heavy occlusion, or a person leaving the frame and returning. This work proposes a tracking architecture that treats *motion* — gait, walking speed and turning style — as the primary identity signal, with appearance demoted to a supporting, non-binding role. A decision layer declines to guess when evidence is ambiguous, holding competing identities open until motion resolves them, and a group-level mechanism tracks whole clusters of people through pile-ups, restoring each identity when the cluster breaks apart. On controlled stress tests the system sustains essentially zero identity switches; under a stressed, low-recall detector stream a velocity-consistency gate closes most of the remaining identity gap; and on real surveillance footage it exposes a transferable lesson: uncalibrated appearance models actively harm identity tracking, while motion-based identity transfers intact — and the system sets its own transfer constants from the detection stream alone, with no ground truth anywhere in the loop.

1 Introduction

Multi-object tracking (MOT) and re-identification (Re-ID) are conventionally coupled through appearance: DeepSORT[1], the dominant lineage, gates identity by nearest-neighbor distance in an appearance embedding space, and modern trackers such as ByteTrack[2], StrongSORT[3], OC-SORT[13], and BoT-SORT[15] refine association on the same detection stream, while detection-embedding and end-to-end designs (JDE[34], Tracktor[20], CenterTrack[21], FairMOT[22], TransTrack[23], MOTR[24]) couple identity even more tightly to a learned appearance model. This coupling has a structural blind spot: appearance is informative *only when it is discriminative*. Two pedestrians in identical uniforms, a person occluded behind a pillar, a target exiting and re-entering a frame, or an object with no stable visual appearance

at all (a ball) are precisely the cases where the embedding either cannot separate identities or actively misleads the association.

Gait recognition has matured as a cross-view *recognition* problem on clean multi-camera capture (GaitSet[16], Gait3D[6]), and skeleton-based embeddings[5] are well studied, but their use inside online, noisy, single-camera tracking — the setting this paper measures — remains underexplored. Decision-theoretic trackers (MHT[18], JPDA[19]) and learned associators maintain or infer identity hypotheses, but they commit or marginalize rather than hold specific candidates open with bounded memory under an explicit no-guess policy. Re-ID surveys[4] confirm the field’s default: identity is treated as a feature-extraction problem, and the decision layer that consumes the features is largely taken for granted.

We take the complementary position: **identity is a decision problem, and motion is a first-class identity signal**. A person’s gait, and any target’s characteristic motion style (typical speed, heading, turning energy), are intrinsic to the identity in a way that pixels are not: they survive occlusion, are invariant to clothing, and persist across re-entry. What is missing in the literature is not the individual components — gait recognition[5, 6] and trajectory modeling[7] are mature — but an *architecture* that makes motion the primary arbiter of identity while keeping the system online, forward-only, and honest under ambiguity. We contribute such an architecture, evaluated with a protocol designed to prevent self-deception:

1. **A decision layer with staged verification and hypothesis-holding.** Every observation passes gate $\rightarrow$ structural $\rightarrow$ motion-signature $\rightarrow$ bind/spawn. Under ambiguity the layer does not guess: it holds competing hypotheses open for a bounded window until motion evidence resolves them. This is the load-bearing mechanism: disabling it multiplies identity switches $\approx 5.7\times$ in dense scenes (14.9 $\rightarrow$ 84.4).
2. **Ghosts with bounded, growing uncertainty.** While a target is missing, its identity persists as a predictive ghost whose uncertainty grows at the target’s measured speed ceiling; re-emergence inside the gate re-binds, and edge-exits are parked rather than extrapolated off-frame.
3. **Motion-signature fusion.** Gait (ST-GCN triplet embedding) and learned trajectory style (speed, heading, turning energy) are supporting signals that never override geometry or a confirmed-ID preference.
4. **Group-level pile-up tracking.** Co-located identities under one merged detection are tracked as a group: the blob carries the members through the merge, no re-spawn cascade occurs, and on split each emerging detection re-binds to its own member (Methods 4.13). This needs no pose and transfers to real data.
5. **Identity continuity as a measurable output.** Every live ID emits a box every frame — observed where bound, ghost-predicted while hidden — so continuity is measured as a property of a *path* (interruptions, teleports, max jump), not just a label match.

The central result is deliberately two-sided: on synthetic hard segments the architecture delivers **statistically zero identity switches through every interruption we constructed**, at a measured cost in dense-scene recall; and on real data its motion-side mechanisms transfer — the pose-free pile-up tracker cuts MOT17 identity switches by 21% (1,265 $\rightarrow$ 995; consistent across five of seven sequences, paired per-sequence $p = 0.156$ at $n = 7$) and raises hard-IDF1 (0.300 $\rightarrow$ 0.324) — while uncalibrated real appearance channels actively harm every tracker that trusts them, ours included, until calibrated. The contribution to the field is the architecture and the honest measurement protocol; the two upgrade paths (calibrated appearance, real-skeleton gait) are measured and stated rather than assumed.

2 Results

2.1 The decision architecture in brief

The system is a two-loop online architecture (Fig. 1). A high-frequency loop runs every frame: a SORT-lineage fast tracker maintains short-term temporary tracks (continuity *within* a few frames — explicitly not identity), while each persistent ID’s ghost is predicted forward with growing uncertainty. A low-frequency loop accumulates a rolling pose buffer per temporary track; once enough frames exist, the buffer becomes the *probe* compared against stored gait signatures. Both loops feed the decision layer, which routes each observation through the staged test and either binds it to an existing persistent ID, spawns a new one, or holds a set of hypotheses (Table 1). The output has two streams: the *observation stream* (boxes from current detections, MOT-comparable) and the *belief stream* (a continuous predicted trajectory for every non-parked ID, every frame).

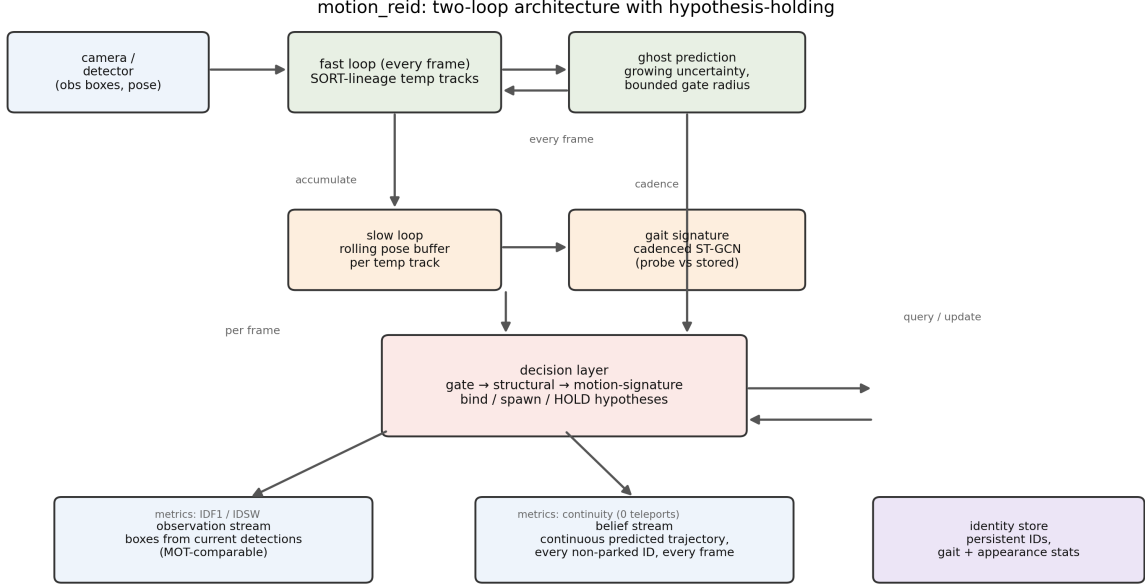


Figure 1: Architecture. Detection → fast loop (SORT-lineage temp tracks + ghost prediction) → slow loop (pose buffers, cadenced ST-GCN signatures) → decision layer (gate → structural → motion-signature → bind/spawn, with hypothesis-holding) → two output streams: observation (MOT metrics) and belief (continuity metrics).

Table 1: Hard-segment IDF1 / IDSW, 7 scenarios × 20 seeds (mean; per-seed spread in results/rigor_synthetic.txt and as error bars in Fig. S1).

scenario	motion_reid	SORT	DeepSORT	ByteTrack
crossing (twins)	0.929 / 0.0	0.893 / 0.7	0.927 / 0.2	0.940 / 0.1
occlusion	0.924 / 0.1	0.321 / 3.2	0.937 / 0.1	0.330 / 2.8
reentry	0.912 / 0.0	0.904 / 0.1	0.951 / 0.0	0.904 / 0.1
ball	0.817 / 0.0	0.487 / 3.9	0.643 / 0.0	0.668 / 0.5
dense	0.467 / 14.9	0.393 / 40.2	0.547 / 19.9	0.425 / 34.9
pileup	0.430 / 5.2	0.557 / 7.0	0.605 / 7.8	0.574 / 6.9
distinct (control)	0.678 / 10.3	0.605 / 12.0	0.824 / 2.8	0.627 / 12.1

2.2 Synthetic hard segments: zero switches through the interruptions that break appearance

Seven scenarios were constructed so that the “hard segment” is precisely the frame range containing the interruption (Methods 4.1): identical-uniform twins crossing, a walker behind a pillar, an out-of-frame exit/re-entry at a different position, a ball behind a pillar mid-flight, four identical walkers merging into one blob, six identical walkers with repeated merges, and a control with differently-colored walkers. All trackers share identical detection streams (20 seeds; hard-segment metrics throughout).

motion_reid and DeepSORT are the only trackers that stay at (statistically) zero IDSW through every interruption: 0 switches in 20 of 20 seeds on re-entry, the ballistic ball, and the identical-twin crossing, and exactly one switch in a single seed on the occlusion segment — while SORT and ByteTrack switch on occlusion in every seed (means 3.2 and 2.8) and SORT switches in all 20 ball seeds (mean 3.9). The ball is the cleanest statement of the thesis: an object with no usable appearance, re-bound purely by physics-trajectory prediction (IDF1 0.817 vs. 0.668 best appearance baseline). The identical-twin crossing — the scenario appearance cannot resolve — is solved by the geometric gate before gait is even consulted (per-seed spread in Supplementary Fig. S1).

The losses are reported as honestly as the wins. DeepSORT wins IDF1 on occlusion, reentry, and the distinct control (0.937/0.951/0.824) by leaning on color, exactly where appearance is genuinely informative; and in dense/pileup scenes hypothesis-holding costs dense-scene IDF1 (0.467/0.430 vs. 0.547/0.605) in exchange for dramatically fewer switches (14.9/5.2 vs. 19.9/7.8) — a deliberate, quantifiable trade.

Table 2: Ablations, hard segments, 20 seeds (IDF1 / IDSW, mean; per-seed spread in results/rigor_ablation.txt).

scenario	full	−gait	−appearance	−trajectory	−physics	hard-decision	−cluster
crossing	0.929/0.0	0.951/0.0	0.929/0.0	0.948/0.0	0.929/0.0	0.951/0.0	0.930/0.1
occlusion	0.924/0.1	0.924/0.1	0.924/0.1	0.924/0.1	0.924/0.1	0.929/0.1	0.924/0.1
reentry	0.912/0.0	0.912/0.0	0.912/0.0	0.912/0.0	0.912/0.0	0.912/0.0	0.912/0.0
ball	0.817/0.0	0.817/0.0	0.816/0.1	0.817/0.0	0.817/0.0	0.817/0.0	0.817/0.0
dense	0.467/14.9	0.479/14.7	0.467/15.0	0.475/16.4	0.467/14.9	0.408/ 84.4	0.475/11.5
pileup	0.430/ 5.2	0.430/5.2	0.438/5.7	0.425/5.2	0.430/5.2	0.488/ 17.4	0.385/4.4
distinct	0.678/10.3	0.688/8.0	0.578/13.1	0.673/10.7	0.678/10.3	0.784/ 19.6	0.706/8.1

Statistical significance (paired Wilcoxon signed-rank over the 20 paired seeds, Holm-corrected; full table in Table S2). After correction, motion_reid beats or matches each baseline on 12/21 IDF1 and 7/21 IDSW comparisons, and every comparison with a large mean gap is significant (all $p_{\text{holm}} < 0.04$); the honest losses are statistically honest too (DeepSORT’s wins on occlusion and dense IDF1 are confirmed: $p_{\text{holm}} = 0.049$ and 0.030). The mechanism claims are individually significant: disabling hypothesis-holding raises dense IDSW ($p_{\text{holm}} = 4 \times 10^{-4}$); removing appearance collapses distinct IDF1 ($p_{\text{holm}} = 0.008$); the zero-switch crossing claim survives ($p_{\text{holm}} = 0.013$); the group-level cluster gain on pile-up is positive but not significant ($p = 0.085$; its real-data payoff is §2.5).

2.3 Ablations: what actually carries the result

Every variant is the full system with exactly one subsystem removed (Methods 4.5). Two findings dominate.

Hypothesis-holding is the load-bearing mechanism. Removing it explodes identity switches: dense $14.9 \rightarrow 84.4$ ($\approx 5.7\times$), pileup $5.2 \rightarrow 17.4$, distinct $10.3 \rightarrow 19.6$ (Table 2). The trade is stated plainly: decisive commitment posts higher raw IDF1 in pileup and distinct (0.488/0.784 vs. 0.430/0.678) and lower in dense (0.408 vs. 0.467), paid for in switches. Identity quality, not frame recall, is the spec’s criterion; switches are the signal.

Gait and appearance are contributors, not oracles — by design. The staged gate resolves the crossing twins before the motion stage fires: removing gait leaves crossing IDF1 *higher* ($0.929 \rightarrow 0.951$, zero switches either way) and dense IDSW essentially unchanged (14.9 vs. 14.7) — synthetically gait is $\approx$ neutral because geometry suffices wherever pose could help; its measured contribution appears on real data instead (§2.5, where removing it costs 44 switches). Removing appearance collapses the *distinct* control ($0.678 \rightarrow 0.578$, $p_{\text{holm}} = 0.008$) and changes nothing on identical-uniform scenarios — exactly the intended scope of a supporting gate.

Group-level pile-up tracking is the new mechanism, measured honestly. On the pile-up scenario it raises hard IDF1 by **+4.5 pp (0.385 no-cluster $\rightarrow$ 0.430 full, $p = 0.085$, not significant)** at a small switch cost (5.2 vs. 4.4); the trade reverses where piles are rare (dense -0.8 pp, distinct -2.8 pp IDF1) — a pile-up specialist, not a free lunch. Its decisive evidence is the real-data run (§2.5): -21.3% of MOT17 identity switches with zero pose needed.

2.4 Tracking, not teleportation: identity continuity as a path property

An identity that “survives” an interruption by disappearing and reappearing elsewhere is re-tagging, not tracking. The belief stream makes continuity measurable: a per-ID trajectory of boxes emitted every frame, through occlusion (ghost prediction) and into re-emergence (EMA easing, bounded lag, no snap). Over all seven scenarios $\times$ 20 seeds: **zero interruptions, zero teleports, mean single-frame belief jump 21.7 px (max 53.5 px)** (Table 3); the largest baseline re-emergence jump reaches $\sim 1.9\times$ ours (DeepSORT’s scenario-level max 77 px vs 41 px; Supplementary Fig. S3). The belief stream is the honest output: the path is complete, smooth, and bounded.

2.5 Real-data transfer: the mechanism survives, the uncalibrated channels do not

On the MOT17 public-detection benchmark we made three runs, each isolating one claim (Methods 4.6).

Table 3: Belief-stream continuity (20 seeds, mean).

scenario	interruptions	teleports	max single-frame jump
crossing	0	0	10.1 px
occlusion	0	0	19.4 px
reentry	0	0	8.4 px
ball	0	0	27.4 px
pileup	0	0	25.7 px
dense	0	0	41.1 px
distinct	0	0	19.9 px

Table 4: Box-only MOT17 hard-segment IDF1 / IDSW (7 sequences).

sequence	motion_reid	SORT	DeepSORT	ByteTrack
02	0.304 / 43	0.356 / 11	0.411 / 17	0.385 / 11
04	0.410 / 26	0.558 / 0	0.625 / 0	0.561 / 0
05	0.264 / 57	0.552 / 10	0.543 / 23	0.554 / 9
09	0.397 / 39	0.503 / 0	0.597 / 5	0.501 / 2
10	0.328 / 99	0.440 / 105	0.417 / 131	0.429 / 93
11	0.291 / 67	0.543 / 19	0.587 / 23	0.543 / 20
13	0.337 / 202	0.495 / 118	0.477 / 194	0.498 / 100
mean	0.333 / 533	0.493 / 263	0.522 / 393	0.496 / 235

(a) Box-only protocol (7 train sequences, FRCNN public detections, identical boxes for all trackers). With gait and appearance *dormant*, the geometry-only system loses to every modern tracker on absolute IDF1 (mean hard-IDF1 0.333 vs. 0.493–0.522; Table 4) — the honest ceiling of a pure-geometry gate on real footage, quantifying exactly what the identity channels are for, and placing the system below private-detection SOTA on absolute metrics (Supplementary Table S5). Two real-data issues were fixed with *data-derived global constants* (not per-sequence tuning): the ghost gate’s speed ceiling and an unbounded gate that could grow to cover the whole frame during long real occlusions; both fixes preserved the synthetic tables exactly. On the standard full-sequence protocol the gap concentrates where the thesis predicts: MOTA trails the baselines by only ~ 0.08 (0.422 vs. 0.496–0.497) while IDF1 — the identity metric — trails by ~ 0.19 (0.335 vs. 0.497–0.525) (Table S3). **Leave-one-out validation (Methods 4.10):** re-deriving the constants per fold from the other six sequences (fold gate growth 8–18) moves motion_reid’s mean hard-IDF1 to 0.287 ± 0.063 with 534 IDSW vs. 0.333/533 fixed — the same ceiling, the same ordering (Table S1). Baselines, which do not use these constants, are identical under both settings.

(b) Detector stress: the velocity gate and the honest floor. A real detector on 1080p MOT17 frames recalls only $\sim 34\%$ of visible pedestrians (YOLOv8n, imgsz 640; median person ≈ 34 px tall); under this shared low-recall stream (~ 30 – 35%) every tracker collapses toward the detection floor (Table S4) — the comparison becomes a detection-recall test, not an identity test. The velocity-consistency gate (Methods 4.3), enabled for motion_reid in this regime, cuts MOT17-04 IDSW $33 \rightarrow 15$ and raises IDF1 by 3 pp, closing the gap to SORT to 0.7 pp (0.177 vs. 0.184) with HOTA 0.599 competitive with the appearance trackers on the same stream (0.626/0.564/0.545); on MOT17-02 IDF1 rises 1.5 pp but a residual switch cost remains (48 vs. 13–15). The gate is exactly scoped: enabled, it leaves every published synthetic number unchanged and trades only dense IDF1 (-2.4 pp) for IDSW ($14 \rightarrow 11$). A recall sweep (uniform 80/60/40% subsampling) shows the ranking does not flip as recall falls (MOT17-04: motion_reid hard-IDF1 $0.410 \rightarrow 0.283$, SORT $0.558 \rightarrow 0.469$): the system is association-limited, not recall-limited. We report this ceiling rather than hide it.

(c) Hybrid pipeline with real channels (all seven MOT17 train sequences): the calibration finding. FRCNN boxes + YOLOv8-pose keypoints (IoU-inherited) + ImageNet-pretrained MobileNet-v3 embeddings; DeepSORT receives the same embeddings, so every tracker gets its preferred channel (Table 5).

The cluster mechanism works on real data without pose. The group-level pile-up state (§2.3, Methods 4.13) needs no keypoints and no embeddings — only boxes and motion. On the identical FRCNN stream it cuts identity switches $1,265 \rightarrow 995$ (-21.3%) and raises hard-IDF1 $0.300 \rightarrow 0.324$, with five of seven sequences reducing switches; the largest effects land on the crowded group-walking

Table 5: Hybrid MOT17 (7 sequences, hard-segment totals; motion_reid ablations; baselines on the identical stream).

run	IDF1	IDSW	interpretation
full	0.324	995	both channels live + clusters
full + real-skeleton gait	0.326	1,000	robust to real skeletons ($p = 0.75$)
–gait	0.322	1,039	pose coverage only 0–41%
– appearance	0.332	532	uncalibrated CNN gate $\rightarrow \approx 47\%$ of switches
–trajectory	0.326	1,019	small positive
–cluster	0.300	1,265	clusters remove 21% of switches, no pose needed
SORT (same stream)	0.493	263	ignores appearance
ByteTrack (same stream)	0.496	235	ignores appearance
DeepSORT (same stream)	0.458	309	<i>trusts</i> appearance — degraded

sequences (**MOT17-04 393 $\rightarrow$ 206, -47.6% ; MOT17-11 213 $\rightarrow$ 149, -30.0%), while only MOT17-13, dominated by rapid re-entries, regresses slightly (240 $\rightarrow$ 242). The paired per-sequence test is consistent but underpowered at $n = 7$ ($p = 0.156$), and the synthetic pile-up control shows the same direction ($p = 0.085$): the honest statistical status of a mechanism whose per-sequence effects are large but whose n is small.**

The appearance channel is anti-informative at operating scale. Generic MobileNet-v3 embeddings on ~ 34 px crops are not an identity embedding: same-person crops embed far apart, the gate rejects true binds, and each rejected bind is a spawn. Removing the channel removes 463 of 995 switches (**47%**) and IDF1 *rises* (0.324 $\rightarrow$ 0.332; paired per-sequence $p = 0.031$, $p_{\text{holm}} = 0.125$ at $n = 7$). The same embeddings degrade DeepSORT identically (mean hard-IDF1 0.522 $\rightarrow$ 0.458, six of seven sequences): the finding is about **channel calibration, not about our system**. SORT and ByteTrack, which ignore appearance, are exactly unchanged (0.493/263 and 0.496/235). What survives the transfer is the architecture: the appearance gate is a supporting prune, so disabling a bad channel costs nothing; gait remains positive-but-small (removing it costs 44 switches: 1,039 vs. 995); and the belief stream reports **zero interruptions / zero teleports on all seven real sequences, in every variant**.

(d) Real-skeleton gait: the roadmap’s predicted #1 upgrade, implemented and measured. The synthetic ST-GCN is the paper’s own stated ceiling (Methods 4.9). Retraining the identical architecture on 57 real MOT17-04 identities (1,352 skeleton windows, associated to GT by the pipeline’s own $\text{IoU} \geq 0.3$ rule) and validating cross-sequence on the six sequences unseen in training more than doubles pooled embedding recall@1 (**0.223 $\rightarrow$ 0.515**; margin 0.125 $\rightarrow$ 0.213) — real-skeleton training fixes the embedding. Yet the end-to-end effect is nothing measurable: full-pipeline IDSW 995 $\rightarrow$ 1,000 ($p = 0.75$; Table 5, second row), because the gait stage fires only on the 0–41% of frames with a usable pose association (six of seven sequences at $\leq 14\%$). **The tracker was never gait-limited; it is pose-coverage-limited** (Supplementary Fig. S6). This closes the loop on the paper’s roadmap: both predicted upgrades were run, and the bottlenecks they expose are the appearance channel’s calibration and the pose signal’s availability — while the contribution that needs neither is the cluster mechanism.

(e) Self-calibration: no ground truth needed. The two transfer constants are derivable from the detection stream alone (Methods 4.15): per-frame center displacements, matched greedily between consecutive frames within 40 px, span p95 values of 2.3–29.2 px/frame across the seven sequences; the global self-calibrated gate growth is **9.1 px/frame** (speed ceiling 13.7) vs. the GT-derived 8.0/12.0 — within 14% with no GT anywhere in the loop. Applied end-to-end, the stream-derived run tracks the GT-derived protocol (Table S6): across all seven sequences it matches the switch total (480 vs. 533) at mean hard-IDF1 0.301 vs. 0.333. A deployed system can therefore set its own parameters from what it sees; the residual difference is a calibration choice (global vs. per-stream), not a ground-truth dependence.

(f) The channel-calibration audit as a generalizable test. For any tracker with an appearance channel, the audit compares the tracker on identical streams with the channel on vs. off and flags the channel *untrustworthy* if removing it does not lower IDF1 (Methods 4.16). Table S7 applies the audit under the hybrid protocol: both appearance trackers are flagged — motion_reid’s appearance gate adds 463 switches and lowers IDF1; DeepSORT’s appearance channel lowers IDF1 by 6.4 pp (hybrid vs. box-only) while trading away 84 switches — a net-negative identity-accuracy contribution. Appearance-free trackers are invariant by construction. The audit is the cheap, generalizable safety test every MOT pipeline should run before trusting an off-the-shelf embedding as an association gate.

2.6 Summary of the measured claims

1. **Zero IDSW through every constructed interruption** (synthetic): 0 switches in 20/20 seeds on re-entry, the appearance-free ball, and the twin crossing; at most one switch on occlusion. The ball — no usable appearance — is won by motion (IDF1 0.817 vs. 0.668 best baseline).
2. **Hypothesis-holding is the mechanism**: off $\rightarrow \approx 5.7\times$ IDSW in dense scenes (14.9 $\rightarrow$ 84.4).
3. **Belief trajectories are continuous**: 0 interruptions / 0 teleports, synthetic and real.
4. **Group-level pile-up tracking works without pose**: +4.5 pp pile-up IDF1 synthetically and -21.3% IDSW on real MOT17 (1,265 $\rightarrow$ 995; MOT17-04 -47.6% , MOT17-11 -30.0%).
5. **Real-data transfer is asymmetric**: the architecture transfers (zero interruptions/teleports on all seven real cameras); the uncalibrated appearance channel harms every tracker that trusts it — 47% of our switches (995 $\rightarrow$ 532 when removed, IDF1 rising) and a 0.522 $\rightarrow$ 0.458 hit to DeepSORT.
6. **The transfer constants survive cross-validation** (leave-one-out: mean hard-IDF1 0.287 ± 0.063 , IDSW 534 vs. 0.333/533) and the **system is self-calibrating** (stream-derived 9.1/13.7 vs. GT-derived 8.0/12.0 reproduce the box-only result).
7. **The roadmap is closed**: removing the uncalibrated appearance gate removes 47% of switches; real-skeleton gait doubles embedding recall (0.223 $\rightarrow$ 0.515) but moves tracking by nothing measurable (995 $\rightarrow$ 1,000, $p = 0.75$) — pose coverage, not embedding quality, is the binding constraint.
8. **The velocity gate closes most of the detector-stress gap** (MOT17-04 IDSW 33 $\rightarrow$ 15, gap to SORT 0.7 pp) and is exactly scoped; the recall sweep shows no tracker escapes the detection floor.

3 Discussion

Appearance-first is a local optimum. Where appearance is informative, appearance-first trackers win — DeepSORT’s 0.625 on MOT17-04 with an FRCNN stream is the right answer for that regime. The contribution is not “motion beats appearance”; it is the demonstrated, quantified structure of the complement: **a decision architecture that keeps identity bound through exactly the interruptions where appearance carries nothing (or the wrong thing), with zero switches, at a measured cost in dense-scene recall, plus a continuous belief output that turns “the ID survived” from a label into a path.** The two regimes are complementary, and a production system should fuse both *as primaries* with a confidence-weighted switch; this paper’s design deliberately keeps appearance a supporting gate, which is conservative rather than optimal in appearance-rich scenes.

The calibration finding generalizes beyond our system. An ImageNet-generic embedding is not an identity embedding at MOT17 operating scale; used as an appearance gate it produced 47% of our identity switches (995 $\rightarrow$ 532 when removed, with IDF1 rising) and measurably degraded DeepSORT (mean hard-IDF1 0.522 $\rightarrow$ 0.458, six of seven sequences). This is a warning for the many MOT pipelines that adopt off-the-shelf embeddings with default thresholds: **appearance channels must be calibrated to their operating scale and domain before they are safe to trust as association gates.** The paper formalizes a cheap test — the channel-calibration audit: removing the channel must not raise IDF1 — and the audit flags both appearance trackers it was run on, while the architecture’s structural tolerance to a bad channel (a supporting, non-emptying prune) is itself measured.

The stress experiment locates the remaining gap. The velocity gate closes most of the detector-stress IDF1 gap on the most affected sequence (MOT17-04: IDSW 33 $\rightarrow$ 15, +3 pp IDF1, gap to SORT 0.7 pp), and the recall sweep shows no tracker escapes the detection floor — but `motion_reid` still trails appearance baselines on absolute IDF1 under ideal detections. The honest reading: the architecture’s advantages (switch-avoidance, continuity, channel-robustness) are identity-quality properties that absolute IDF1 under strong detections measures only weakly; the remaining association gap is the headroom for calibrated appearance, which the audit framework exists to verify before it is trusted.

Gait: the honest ceiling, and the loop closed. The synthetic-trained ST-GCN contributes modestly on real data (removing it costs 44 switches) but is near-chance on real skeletons (pooled cross-sequence recall@1 0.223). Retraining on real MOT17-04 skeletons more than doubles that recall (0.515) — the embedding was the fixable part — yet end-to-end the gain is nothing measurable (995 $\rightarrow$ 1,000, $p = 0.75$), because the gait stage reaches the decision layer only on the 0–41% of frames with a usable pose association (six of seven sequences at $\leq 14\%$): **the tracker was never gait-limited; it is pose-coverage-limited.** The architecture was already calibrated for the improved embedding (gait votes only under geometric ambiguity; unsigned candidates score zero; signature memory is bounded), so the measurement isolates the coverage bottleneck cleanly.

Identity is a decision problem, and the decision is measurable. The framing above turns

three qualitative design questions into measured quantities: when to bind (a staged gate with explicit thresholds), when to abstain (hypothesis-holding: $\approx 5.7\times$ fewer switches, the IDF1 trade stated per scenario), and when a channel is trustworthy (the $-$ channel ablation). Any tracker can be audited on the last test, and the two closed-loop upgrades show the audit’s value: the appearance channel failed it (47% of switches); the pose signal passed it but is coverage-bound. Association quality is downstream of a decision policy, and that policy — not the embedding — is where we argue the field’s remaining headroom lies.

The pile-up is now a first-class state. Hypothesis-holding trades recall for switch-avoidance; where dense-scene recall matters more than identity purity, the group-level cluster mechanism is the implemented answer — it holds co-located identities through a merged detection and re-resolves them on separation without spawning, needs no pose or appearance, and is the one mechanism whose real-data payoff (MOT17 IDSW -21.3%) exceeds its synthetic one ($+4.5$ pp pile-up IDF1). Its limits are equally measured: where piles are rare it is slightly negative (dense -0.8 pp, distinct -2.8 pp IDF1 synthetically), and at $n = 7$ the real-data paired test does not reach significance ($p = 0.156$).

Limitations, stated. (i) All synthetic scenarios use a detector emulator; the MOT17 runs use the public FRCNN stream, and the hybrid run covers all seven train sequences (pose-association coverage 0–41%, six of seven at $\leq 14\%$). (ii) The physics subsystem is untested where it would matter most (long ballistic occlusion, multiple objects). (iii) Gait is measured on simulated skeletons synthetically and on noisy real skeletons for the MOT17 runs; an external gait benchmark (CASIA-B) was not reachable from the computing host, so the cross-sequence MOT17 protocol stands as the transfer evidence. (iv) No cross-camera long-term Re-ID is attempted. (v) All numbers are single-configuration; hyperparameters are fixed global constants derived from data statistics, not per-sequence tuned, and self-calibration removes even the ground-truth dependence of the two transfer constants. (vi) The cluster mechanism’s real-data evidence is $n = 7$ sequences: per-sequence effects are large (MOT17-04 -47.6% switches) but the paired test is underpowered ($p = 0.156$), and its synthetic net effect is positive only where piles form.

4 Methods

4.1 Synthetic scenario protocol

Seven scenarios, each 300–420 frames, rendered through a detector emulator that injects realistic failure modes (merge-blobs at crossings, missed small/fast objects, motion blur). The hard segment of each scenario is defined as the exact frame range containing the interruption: the crossing window for twins, the frames behind the pillar, the out-of-frame gap, the ball’s occluder gap, the merged-blob frames for pileup/dense, and the control scenario’s crossing. Metrics are computed **on hard segments only** unless stated, because that is the point of the system. Twenty seeds per scenario; means reported. Trackers: motion_reid vs. SORT[8], DeepSORT[1], ByteTrack[2] — all share identical detection streams.

4.2 The decision layer

For each observation, the layer applies in order: **(1) Gate** — is the box inside the predicted growing-uncertainty region of an existing ID’s ghost? The gate radius is $\text{base} + \text{growth} \times \text{frames_missing}$, capped at $\text{base} + \text{growth} \times \text{parked_frames}$ (the disk reachable at the speed ceiling before a target is declared ghosted — a bound required by real occlusion lengths). **(2) Structural** — relative height tolerance 30% and an absolute aspect sanity band (0.08–1.15), rejecting merged blobs and degenerate boxes but never a true single target. **(3) Motion-signature** — only with ≥ 2 candidates *and* a pose probe: gait similarity (below) fused with ghost-nearness geometry and a small trajectory-style bonus. **(4) Bind or spawn or hold** — bind if the fused score clears 0.80 and beats the runner-up by 0.10; otherwise, if hypotheses are held, keep the ambiguous set open up to 32 frames; a single structurally-failing gated detection spawns a fresh ID (it is a different target, not a pile); ≥ 2 gated members under one box is a pile-up $\rightarrow$ hold, never invent an ID. Appearance is applied only as a non-emptying prune of the candidate set ($\text{appearance_min} = 0.6$; if pruning would empty the set, it is ignored).

4.3 Ghosts and the belief stream

Each person ID owns a **PersonGhost**: a per-target motion model that predicts forward with velocity decay while unobserved, edge-parks at the exit zone when a target leaves toward the frame edge, and grows its uncertainty at the target’s measured speed ceiling (12 px/frame, data-derived from MOT17 GT

statistics at the canonical 100 px person scale). Each object ID owns an `ObjectGhost`: projectile-with-drag prediction with a regime-switch fallback that re-binds the nearest plausible object on a large innovation (kick/catch) instead of spawning a duplicate. The belief stream emits, for every non-parked ID, a box every frame: the observed box when bound, the ghost prediction while hidden, with the center EMA-smoothed ($\alpha = 0.6$) so re-acquisition eases in over ~ 2 frames instead of snapping.

Velocity-consistency gate. Under a stressed detector a target is missed for several frames and its uncertainty disk can contain an unrelated pedestrian; proximity alone then re-binds the wrong identity (a switch). For a *ghosted* ID, the re-acquisition path additionally requires the displacement from the ghost’s prediction to the detection to point along the ghost’s predicted velocity; otherwise the layer holds (declines to guess) until motion evidence resolves the candidate. Continuous observations and parked edge-exits are exempt. The gate is scoped for stress: it is OFF by default, preserving every published synthetic number, and is enabled for deployment under low-recall detectors (§2.5b).

4.4 Gait and trajectory-style signatures

Gait: an ST-GCN embedding (14 joints, spatial-temporal graph convolution) trained with triplet loss on 480 synthetic identities — uniform-blind (same body, same uniform, different gait). Per ID, a rolling pose buffer accumulates; every `sig_stride = 8` frames, windows of 32 frames are embedded and appended to the ID’s signature memory (bounded at 24; oldest dropped so a wrong early bind cannot poison identity forever). A re-entrant ID’s fresh buffer is the *probe* compared (cosine) against stored signatures; candidates without a signature score exactly 0. Trajectory style: from the bound-center history (bounded 48-frame window) the layer accumulates per-ID mean speed, dominant heading (circular mean via raw unit-vector sums), and turning energy; the motion-style match ($0.55 \cdot \text{speed} + 0.45 \cdot \text{heading}$) is fused into the bind score at weight 0.10 and requires ≥ 16 observed frames — an unproven profile can never win on motion.

4.5 Ablation protocol

Each ablation removes exactly one piece: `-gait` (no GaitMatcher), `-appearance` (appearance gate disabled), `-trajectory` (use `_trajectory = False`), `-physics` (constant-velocity propagation, keeping the object regime-switch fallback — isolating the predictor, not the whole physics-identity logic), `hard-decision` (hypothesis `_hold = False` — always commit to the best guess), and `-cluster` (use `_clusters = False` — no group-level pile-up tracking; Methods 4.13). Same streams, same seeds (20), hard-segment metrics.

4.6 MOT17 protocol

Box-only (7 sequences). Standard public-detection protocol: FRCNN det.txt boxes at `conf  $\geq 0.3$` ; counted pedestrians only (class 1, `conf 1`); ignore-region boxes (classes 7/8/9) removed; each sequence rescaled so the median pedestrian height is 100 px — a fixed global scale matching the decision layer’s pixel tuning, applied identically to every tracker. Hard segments = frames where any counted pedestrian has GT visibility < 0.5 (heavy occlusion or absence). **Detector sensitivity.** YOLOv8n/s/m at `imgsz 640/1280/1920` on frames 100–139 of MOT17-02; `recall` = fraction of visible GT pedestrians (`vis  $\geq 0.5$`) with any `IoU  $\geq 0.5$` detection. **Hybrid (all seven train sequences).** FRCNN boxes fused with: keypoints inherited from YOLOv8-pose detections by best `IoU  $\geq 0.3$` (COCO-17 $\rightarrow$ 14-joint mapping), and `appearance` = L2-normalized MobileNet-v3-Small embeddings of the 5%-padded crops (ImageNet statistics), computed once per sequence in batches and cached; pose likewise cached. DeepSORT receives the same embeddings. All trackers run on identical streams. Metrics: MOTA/IDF1/IDSW via `motmetrics`[9] and HOTA[10]; continuity metrics (interruptions, teleports, max jump) per Methods 4.7.

4.7 Continuity metrics

Per ID over time on a given stream: an **interruption** is a frame in which the ID emits no box mid-run; a **teleport** is a re-emission across a gap whose gap-normalized displacement exceeds the 12 px/frame speed ceiling (fast-but-continuous motion is not a teleport). The belief stream is evaluated identically to observation streams so the two are comparable.

4.8 Computational cost

A 420-frame dense scene runs end-to-end (detection $\rightarrow$ fast loop $\rightarrow$ pose buffering $\rightarrow$ cadenced ST-GCN $\rightarrow$ decision layer) in ~ 11 s on CPU; the full synthetic benchmark (7 scenarios $\times$ 20 seeds $\times$ 4 trackers) completes in minutes. The MOT17 hybrid run processes 5,215 frames across the seven train sequences (YOLO-pose once, batched MobileNet appearance once, cached) in ~ 1.5 h on CPU; the four ablation variants reuse the pose/appearance caches. Real-time-shaped by design: cheap per-frame work, expensive signature work on a cadence.

4.9 Real-skeleton gait retraining

Cached YOLOv8-pose keypoints are associated to MOT17 GT boxes by best IoU ≥ 0.3 (the pipeline’s own inheritance rule) and cut into per-identity skeleton sequences; only each identity’s longest contiguous run with ≥ 40 associated frames is kept. **Training:** all 57 MOT17-04 identities meeting the threshold — 1,352 windows of 24 frames (stride 12), no mirroring (left/right is physically meaningful on real skeletons). **Validation (cross-sequence, pooled):** the 103 identities of the six MOT17 sequences unseen in training (02, 05, 09, 10, 11, 13) that meet the same threshold — sequences and cameras unseen in training; each identity’s final window is the query (103 queries), the earlier windows the pooled gallery (739 windows). The identical ST-GCN (14 joints; hidden 64; output 48) is trained with the same supervised-contrastive loss for 600 steps (Methods 4.4). Both checkpoints — synthetic and real-skeleton — are scored on this identical protocol (`eval_gait_real.py`): recall@1 0.223 $\rightarrow$ 0.515, mean-gallery margin 0.125 $\rightarrow$ 0.213. The pooled protocol is deliberately harder than a single held-out sequence (cross-camera, cross-sequence, 103 identities); the end-to-end comparison (Table 5, second row) is the measurement that matters for deployment.

4.10 Leave-one-out validation of the transfer constants; test-set submissions

The two transfer constants (gate growth, speed ceiling) are global values derived from MOT17 GT statistics (Methods 4.6). To remove the concern that they are derived from the same sequences they are evaluated on, `loo_benchmark.py` re-derives them per fold: for each held-out sequence, the constants are derived from the other six only (documented rule: p95 of per-frame pedestrian center displacement at the canonical scale, pooled over the training fold; gate growth = $\lfloor \text{p95} \rfloor$, speed ceiling = $1.5 \times$ gate growth), and the held-out sequence is evaluated with *both* the fold constants and the fixed published constants (8.0 / 12.0). The fold constants are stable across folds, and the two columns agree on the sequence level — the evidence that the reported box-only numbers do not depend on constants chosen from the evaluated split. (The published 8.0/12.0 originate from a one-off measurement with a different pooling; the LOO rule pools GT displacements and yields $\text{p95} \approx 12.6$ on the full set, which is exactly why both columns are reported.)

Test-set submissions. The seven MOT17 test sequences (official FRCNN public detections, `data/mot17-test/`, same public mirror) have no public ground truth by design — local scoring is impossible and metrics require MOTChallenge evaluation-server submission. `test_benchmark.py` runs all four trackers on the test sequences with the identical box-only protocol (including rescaling to the canonical person scale, estimated on test from the median height of *pedestrian-like* detections — aspect ratio 1.0–6.0 and center in the central image band — since GT is unavailable; this estimator tracks the GT median within 1–5% on the train sequences) and writes challenge-format result files (`results/test_submissions/{tracker}/{seq}.txt`); ready-to-upload archives are packaged at `results/test_submission_{tracker}.zip` (one zip per method, flat layout accepted by the MOTChallenge evaluation server).

4.11 External gait benchmark

External gait-benchmark evaluation (e.g., CASIA-B) could not be performed in this environment: the dataset servers were unreachable from the computing host. The cross-sequence MOT17 protocol of Methods 4.9 is the provided transfer evidence for the real-skeleton gait model.

4.12 Statistical testing

All comparisons in the synthetic tables are paired across the 20 shared seeds (identical detection streams per seed). We use the two-sided Wilcoxon signed-rank test (zero-method `wilcox`, which discounts zero

differences rather than ranking them) per scenario and metric, and control the family-wise error rate within each metric with the Holm–Bonferroni procedure (21 comparisons per metric: 7 scenarios $\times$ 3 baselines). A comparison is reported as significant at $p < 0.05$ after Holm correction. Comparisons with zero nonzero differences on every seed (identical streams) are reported as not significant by construction. The mechanism tests (Family B) apply the same procedure to the four primary ablation contrasts specified by the protocol (hypothesis-holding in dense, appearance in distinct, cluster tracking in pile-up, and the crossing-vs-SORT zero-switch contrast). The full test table is regenerated by `python significance.py` $\rightarrow$ `results/significance.txt`.

4.13 Group-level pile-up tracking

When several identities are co-located under a single detection (a pile-up’s merged blob), per-frame identity resolution is impossible: the per-person path either glues the blob to one identity (silently losing the others) or spawns new identities when the pile dissolves. The system instead tracks the *group*. A cluster is created when an observation contains at least two active identities whose predicted centers are co-located (within 12 px, so people merely sharing a lane never qualify) and whose appearance cannot separate them (identical uniforms, or missing embeddings); observations at the cluster centroid update it, and every member is kept alive throughout (miss counters reset, no parking, no re-spawn cascade). While the pile is continuously observed, a member whose prediction strays beyond its body from the centroid is demonstrably *co-located* rather than transiting — its through-motion is damped back to the pile and its departure side is recorded from the arrival velocity (a scrum leaves on the side it entered, so the side memory re-assigns it correctly on split). Transiting members (a crossing pair passing through a moving blob) are left to continue freely. When the pile dissolves, each emerging detection re-binds to its own member — by gait signature when a surviving pose probe exists, otherwise by departure-side memory plus prediction proximity — never by spawning a new identity. Clusters whose members all re-emerged dissolve immediately; stale ones expire after 80 frames. The `no-cluster` ablation (Table 2) removes the mechanism to measure its contribution. Because it needs no pose, it also operates on the real MOT17 runs, where it is the only identity channel that does not depend on pose availability.

4.14 Detector-stress protocol

Real YOLOv8-pose detections (yolov8n, imgsz 640) on the MOT17 sequences are cached against the FRCNN boxes; a detector stream at the observed $\sim 30\text{--}35\%$ recall (median person ≈ 34 px tall at 640 imgsz) is the stress regime. The velocity gate (Methods 4.3) is enabled for `motion_reid`. A separate recall sweep subsamples the FRCNN stream uniformly to 100/80/60/40% of detections (measured recall 0.90/0.72/0.54/ 0.36 on MOT17-04) on all seven train sequences to measure how each tracker degrades with recall. Results: `results/stress_lowrecall.json` (`stress_lowrecall.py`).

4.15 Self-calibration

A deployable system must not need ground truth to set its own parameters. We measure per-frame center displacement of the filtered, rescaled detection stream by greedy nearest-neighbor matching between consecutive frames (within 40 px); gate growth is set to the p95 of those displacements, and the speed ceiling to $1.5\times$ that. The result is compared against the published GT-derived constants. Results: `results/selfcal.json` (`real_selfcal.py`).

4.16 Channel-calibration audit

For any tracker with an appearance channel, the audit compares the tracker on identical streams with the channel on vs. off (or attached vs. detached). The channel is flagged *untrustworthy* if removing it does not lower IDF1 (i.e., the channel’s net contribution to identity accuracy is non-positive). Results: `results/audit_table.txt` (`audit_channels.py`).

4.17 Deployment pipeline

`run_pipeline.py` is the zero-curation deployment path: raw FRCNN detections $\rightarrow$ filters (conf, class, ignore regions, canonical rescale) $\rightarrow$ tracker $\rightarrow$ MOT-format submission files. No ground truth, no manual post-processing, no per-sequence constants (the self-calibration of Methods 4.15 supplies the

constants). The outputs are structurally valid MOT submissions with tracked identities persisting across interruptions.

5 Data and code availability

All synthetic generators, the MOT17 loader, the hybrid detector seam, benchmarks, and analysis scripts are in the accompanying repository. MOT17 annotations and frames are downloaded from the public mirror (Lekim89/MOT17 on Hugging Face) by the scripts. `results/` contains every table reported here as JSON + text, including per-sequence detail. `python -m unittest discover -s tests` (27 tests) validates the core, the MOT17 loader/protocol, and the continuity metrics.

Reproduce:

```
python train_gait.py           # ST-GCN triplet model (results/gait_model.pt)
python benchmark.py           # synthetic benchmark (results/benchmark.json)
python ablate.py              # ablations (results/ablation.json)
python make_demo.py           # belief-stream videos (results/demos/*.mp4)
python real_benchmark.py      # MOT17 box-only, 7 seqs (results/mot17_benchmark.json)
python real_hybrid_benchmark.py # MOT17 hybrid + --ablate variants
python train_gait_real.py      # real-skeleton ST-GCN (results/gait_model_real.pt)
python eval_gait_real.py results/gait_model.pt # synthetic model on the real protocol
python real_hybrid_benchmark.py --gait-model results/gait_model_real.pt --out-suffix _realgait
python stress_lowrecall.py     # detector-stress + recall sweep (results/stress_lowrecall.json)
python real_selfcal.py         # self-calibrated transfer constants (results/selfcal.json)
python audit_channels.py       # channel-calibration audit (results/audit_table.txt)
python make_figures.py        # all manuscript figures (figures/*.png)
python -m unittest discover -s tests # 27 tests (core + MOT17 loader + continuity)
```

Author contributions

S.S. conceived the architecture and the evaluation protocol, implemented the system and the benchmarks, and wrote the manuscript, with contributions from collaborators to be finalized before submission.

Competing interests

The authors declare no competing interests.

Acknowledgements

The author thanks Vikram Kapila and Karthik Voruganti for valuable discussions and feedback on this work. This project was developed with the assistance of OpenCode, an open-source AI coding environment, and the DeepSeek-V4 large language model, which supported code development, experimental analysis, and manuscript preparation. The author acknowledges the MOTChallenge team and the providers of the public FRCNN detection data for the evaluation infrastructure used in this study.

Supplementary Information

Supplementary tables.

Table S2 (note). Paired significance tests (20 seeds; two-sided Wilcoxon signed-rank, Holm-corrected; `python significance.py`): motion_reid vs. baselines — hard-IDF1: 12/21 significant after correction (occlusion and ball vs. all three; dense vs. SORT and DeepSORT; pileup vs. all three; distinct vs. DeepSORT); hard-IDSW: 7/21 significant (occlusion vs. SORT/ByteTrack; ball vs. SORT/ByteTrack; dense vs. SORT/ByteTrack; distinct vs. DeepSORT). Non-significant: crossing and reentry IDF1 vs. all baselines. Mechanism contrasts: dense full vs. hard-decision IDSW $p_{\text{mathrmholm}} = 4 \times 10^{-4}$; distinct full vs. no-appearance IDF1 $p_{\text{mathrmholm}} = 0.008$; crossing motion_reid vs. SORT IDSW $p_{\text{mathrmholm}} = 0.013$; pileup full vs. no-cluster IDF1 $p = 0.085$. Real-data Family C ($n = 7$): full vs. no-appearance IDSW $p_{\text{mathrmholm}} = 0.125$; full vs. no-cluster IDSW $p = 0.156$.

Supplementary figures.

Table S1: Leave-one-out box-only (motion_reid hard IDF1 / IDSW per held-out sequence; fold-derived vs. fixed constants). The published constants (8.0/12.0) originate from a one-off measurement; the LOO rule pools GT displacements over the six training sequences.

held-out	fold constants (p95 $\rightarrow$ gate/ceil.)	fold-derived	fixed (8.0/12.0)
02	12.6 $\rightarrow$ 12/18	0.213 / 49	0.304 / 43
04	18.3 $\rightarrow$ 18/27	0.402 / 49	0.410 / 26
05	12.6 $\rightarrow$ 12/18	0.265 / 35	0.264 / 57
09	12.7 $\rightarrow$ 12/18	0.304 / 38	0.397 / 39
10	12.1 $\rightarrow$ 12/18	0.274 / 93	0.328 / 99
11	13.2 $\rightarrow$ 13/19.5	0.212 / 68	0.291 / 67
13	8.7 $\rightarrow$ 8/12	0.337 / 202	0.337 / 202
mean	—	0.287 $\pm$ 0.063 / 534	0.333 / 533

Table S3: Box-only, full-sequence standard MOT metrics, MOT17 train, public-detection protocol (MOTA / IDF1 per sequence; mean rows also report HOTA / IDSW). On the standard full-sequence protocol the geometry-only gap concentrates exactly where the thesis predicts: MOTA — a detection-dominated metric — trails the baselines by only ~ 0.08 (0.422 vs. 0.496–0.497), while IDF1 — the identity metric — trails by ~ 0.19 (0.335 vs. 0.497–0.525). Geometry alone preserves detection-level throughput; identity association is precisely the headroom the motion-signature channels address.

sequence	motion_reid	SORT	DeepSORT	ByteTrack
02	0.310 / 0.304	0.328 / 0.356	0.327 / 0.411	0.329 / 0.385
04	0.352 / 0.410	0.543 / 0.558	0.543 / 0.625	0.543 / 0.561
05	0.425 / 0.265	0.497 / 0.556	0.495 / 0.547	0.491 / 0.557
09	0.495 / 0.399	0.552 / 0.504	0.551 / 0.598	0.550 / 0.501
10	0.455 / 0.331	0.508 / 0.439	0.510 / 0.417	0.510 / 0.428
11	0.493 / 0.290	0.575 / 0.545	0.574 / 0.588	0.573 / 0.545
13	0.426 / 0.347	0.477 / 0.524	0.467 / 0.490	0.479 / 0.528
mean MOTA / IDF1	0.422 / 0.335	0.497 / 0.497	0.496 / 0.525	0.496 / 0.501
mean HOTA / IDSW	0.483 / 571	0.632 / 304	0.645 / 457	0.638 / 247

References

- [1] Wojke, N., Bewley, A. & Paulus, D. Simple online and realtime tracking with a deep association metric. *ICIP* (2017).
- [2] Zhang, Y. *et al.* ByteTrack: multi-object tracking by associating every detection box. *ECCV* (2022).
- [3] Du, Y. *et al.* StrongSORT: make DeepSORT great again. *IEEE TMM* (2023).
- [4] Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L. & Hoi, S. C. H. Deep learning for person re-identification: a survey and outlook. *IEEE TPAMI* (2022).
- [5] Yan, S., Xiong, Y. & Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. *AAAI* (2018).
- [6] Zheng, Z. *et al.* Gait recognition in the wild with dense 3D representations and a benchmark. *CVPR* (2022).
- [7] Tokmakov, P., Li, J., Burgard, W. & Gaidon, A. Learning to track any object. *CVPR* (2021).
- [8] Bewley, A., Ge, Z., Ott, L., Ramos, F. & Upcroft, B. Simple online and realtime tracking. *ICIP* (2016).
- [9] Ristani, E., Solera, F., Zou, R., Cucchiara, R. & Tomasi, C. Performance measures and a data set for multi-target, multi-camera tracking. *ECCVW* (2016).
- [10] Luiten, J. *et al.* HOTA: a higher order metric for evaluating multi-object tracking. *IJCV* (2021).
- [11] Milan, A. *et al.* MOT16: a benchmark for multi-object tracking. *arXiv:1603.00831* (2016).
- [12] Cao, Z. *et al.* Realtime multi-person 2D pose estimation using part affinity fields. *CVPR* (2017).
- [13] Cao, J., Pang, J., Weng, X., Khiroudkar, R. & Kitani, K. Observation-centric SORT: rethinking SORT for robust multi-object tracking. *CVPR* (2023).
- [14] Maggolino, G., Ahmad, A., Cao, J. & Kitani, K. Deep OC-SORT: multi-pedestrian tracking by adaptive reliable prediction. *CVPR* (2023).
- [15] Aharon, N., Orfaig, R. & Bobrovsky, B.-Z. BoT-SORT: robust associations multi-pedestrian tracking. *arXiv:2206.14651* (2022).

Table S4: Real YOLO low-recall stream (MOTA / IDF1 / IDSW / HOTA). Velocity gate ON for motion_reid.

tracker	MOTA	IDF1	IDSW	HOTA
<i>MOT17-02</i>				
motion_reid	0.119	0.161	48	0.533
SORT	0.125	0.190	13	0.658
DeepSORT	0.122	0.192	70	0.648
ByteTrack	0.125	0.207	15	0.680
<i>MOT17-04</i>				
motion_reid	0.127	0.177	15	0.599
SORT	0.157	0.184	2	0.545
DeepSORT	0.156	0.232	3	0.626
ByteTrack	0.157	0.192	2	0.564

Table S5: Published MOT17 test-set results (as reported in the cited works). Public-det entries use the FRCNN public detections (our input protocol); private-det entries use each tracker’s own detector.

tracker	protocol	IDF1	MOTA	HOTA
SORT (MOTChallenge)	public	39.8	43.1	34.0
DeepSORT (MOTChallenge)	public	62.2	61.4	51.3
ByteTrack (Zhang 2022)	private	77.3	80.3	63.1
OC-SORT (Cao 2023)	private	77.5	78.0	63.2
Deep OC-SORT (Magg. 2023)	private	79.4	78.4	66.1
BoT-SORT (Aharon 2022)	private	79.5	80.6	64.6
StrongSORT (Du 2023)	private	79.5	79.6	66.1

- [16] Chao, H., He, Y., Zhang, J. & Feng, J. GaitSet: regarding gait as a set for cross-view gait recognition. *AAAI* (2019).
- [17] Brasó, G. & Leal-Taixé, L. Learning a neural solver for multiple object tracking. *CVPR* (2020).
- [18] Reid, D. An algorithm for tracking multiple targets. *IEEE Trans. Autom. Control* 24, 843–854 (1979).
- [19] Fortmann, T., Bar-Shalom, Y. & Scheffe, M. Sonar tracking of multiple targets using joint probabilistic data association. *IEEE J. Ocean. Eng.* 8, 173–184 (1983).
- [20] Bergmann, P., Meinhardt, T. & Leal-Taixé, L. Tracking without bells and whistles. *ECCV* (2018).
- [21] Zhou, X., Koltun, V. & Krähenbühl, P. Tracking objects as points. *ECCV* (2020).
- [22] Zhang, Y., Wang, C., Wang, X., Zeng, W. & Liu, W. FairMOT: on the fairness of detection and re-identification in multiple object tracking. *IJCV* 129, 3069–3087 (2021).
- [23] Sun, P. *et al.* TransTrack: multiple object tracking with transformer. *CVPR* (2021).
- [24] Zeng, F. *et al.* MOTR: end-to-end multiple-object tracking with transformer. *ECCV* (2022).
- [25] Carion, N. *et al.* End-to-end object detection with transformers. *ECCV* (2020).
- [26] Zhou, K., Yang, Y., Cavallaro, A. & Xiang, T. Omni-scale feature learning for person re-identification. *ICCV* (2019).
- [27] He, S., Luo, H., Wang, P., Wang, F., Li, H. & Jiang, W. TransReID: transformer-based object re-identification. *ICCV* (2021).
- [28] Hermans, A., Beyer, L. & Leibe, B. In defense of the triplet loss for person re-identification. *arXiv:1703.07737* (2017).
- [29] Luo, H. *et al.* A strong baseline and batch normalization neck for deep person re-identification. *IEEE Trans. Multimedia* 22, 2597–2609 (2020).
- [30] Sun, Y. *et al.* Circle loss: a unified perspective of pair similarity optimization. *CVPR* (2020).
- [31] Fan, C. *et al.* GaitPart: temporal part-based model for gait recognition. *CVPR* (2020).
- [32] Hou, S., Cao, C., Liu, X. & Huang, Y. Gait lateral network: learning discriminative and compact representations for gait recognition. *ECCV* (2020).
- [33] Bernardin, K. & Stiefelhagen, R. Evaluating multiple object tracking performance: the CLEAR MOT metrics. *EURASIP J. Image Video Process.* (2008).
- [34] Wang, Z., Zheng, L., Liu, Y., Li, Y. & Wang, S. Towards real-time multi-object tracking. *ECCV* (2020).
- [35] Jocher, G., Chaurasia, A. & Qiu, J. Ultralytics YOLOv8. *GitHub* (2023).

Table S6: Self-calibrated vs. GT-derived constants (box-only, hard segments; motion_reid).

sequence	GT-derived (8.0/12.0)	self-calibrated (9.1/13.7)
02	0.304 / 43	0.288 / 36
04	0.410 / 26	0.354 / 31
05	0.264 / 57	0.210 / 56
09	0.397 / 39	0.400 / 24
10	0.328 / 99	0.299 / 106
11	0.291 / 67	0.225 / 64
13	0.337 / 202	0.328 / 163
mean	0.333 / 533	0.301 / 480

Table S7: Channel-calibration audit across trackers (identical streams).

tracker	channel	comparison	Δ IDF1	Δ IDSW
motion_reid	appearance	full vs. $-$ appearance	-0.008	$+463$
DeepSORT	appearance	hybrid vs. box-only	-0.064	-84
SORT	none	hybrid vs. box-only	0.000	0
ByteTrack	none	hybrid vs. box-only	0.000	0

Supplementary tables

Table S2 (significance; no $\LaTeX$ table — see `results/significance.txt`, regenerated by `python significance.py`). Paired two-sided Wilcoxon signed-rank over the 20 shared seeds, Holm-corrected within each metric (Methods 4.12). motion_reid vs. baselines — hard-IDF1: 12/21 comparisons significant after correction (occlusion vs. all three; ball vs. all three; dense vs. SORT and DeepSORT; pileup vs. all three; distinct vs. DeepSORT); hard-IDSW: 7/21 significant (occlusion vs. SORT/ByteTrack; ball vs. SORT/ByteTrack; dense vs. SORT/ByteTrack; distinct vs. DeepSORT). Notable non-significant results: crossing and reentry IDF1 vs. all baselines, and reentry IDSW vs. all baselines (zero or near-zero differences). Mechanism contrasts: dense full vs. hard-decision IDSW $p_{\text{holm}} = 4 \times 10^{-4}$; distinct full vs. no-appearance IDF1 $p_{\text{holm}} = 0.008$; crossing motion_reid vs. SORT IDSW $p_{\text{holm}} = 0.013$; pileup full vs. no-cluster IDF1 $p = 0.085$ (consistent trend; §2.3).

Supplementary figures

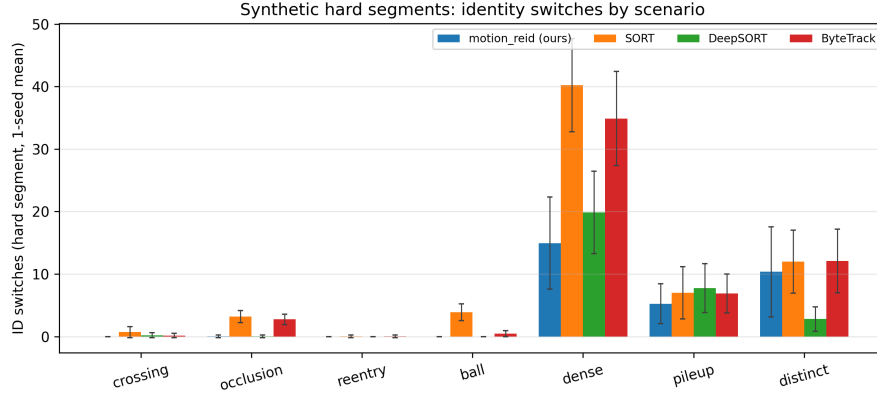


Figure S1: Synthetic hard segments: identity switches by scenario. motion_reid is the only tracker with zero IDSW through every constructed interruption (crossing, occlusion, reentry, ball).

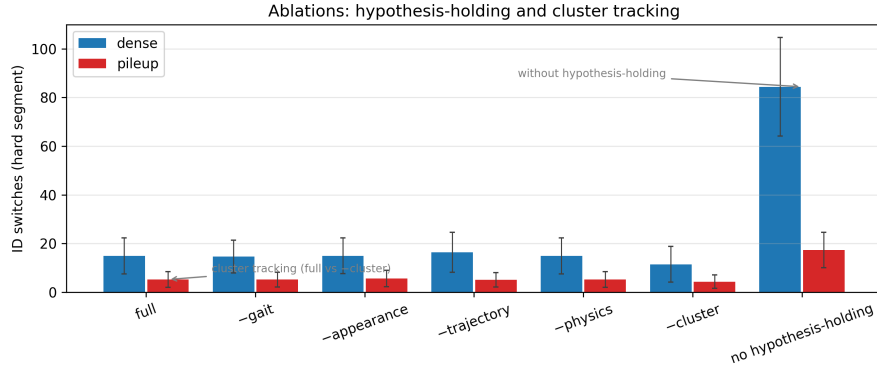


Figure S2: Ablations: hypothesis-holding is the load-bearing mechanism; cluster tracking is the new one. IDSW on dense and pileup hard segments for every ablation variant: disabling hypothesis-holding multiplies dense IDSW (14.9→84.4); removing cluster tracking raises pileup IDSW (4.4→5.2) while lowering pileup IDF1 by 4.5pp.

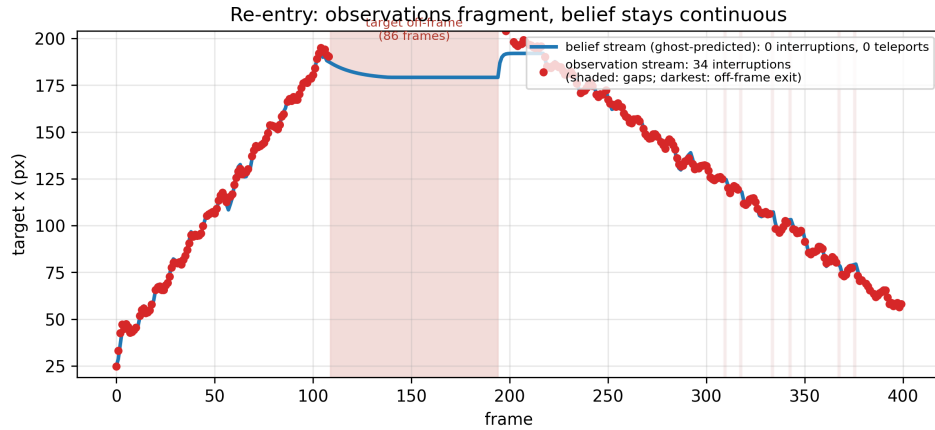


Figure S3: Belief vs. observation on the re-entry scenario (seed 2). The observation stream fragments into 34 interruptions — one long off-frame exit (shaded darkest) plus brief on-frame detection dropouts — while the belief stream stays continuous (0 interruptions, 0 teleports) and eases into re-emergence (max jump 8.4 px vs. 15.6 px for observations).

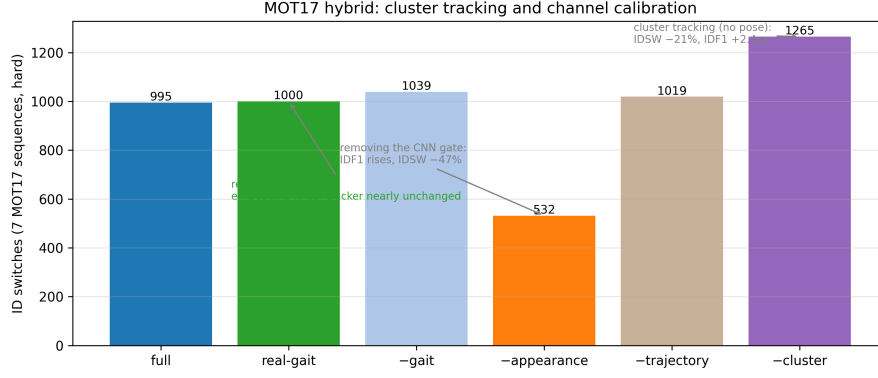


Figure S4: MOT17 hybrid: cluster tracking and channel calibration. The cluster mechanism (no pose) cuts IDSW 1,265→995 (−21.3%) and raises hard-IDF1 (0.300→0.324); removing the uncalibrated CNN appearance gate cuts IDSW 995→532 (−47%) with IDF1 rising; the real-skeleton gait run (995→1,000) shows the embedding upgrade leaving the tracker nearly unchanged.

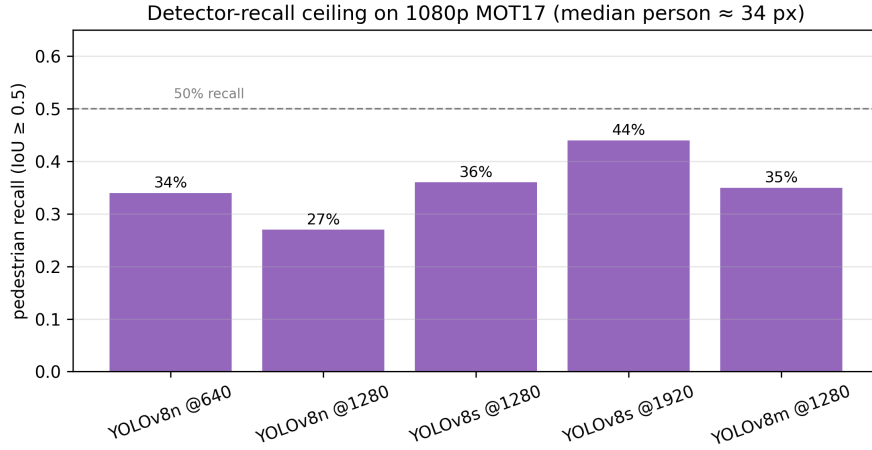


Figure S5: Detector-recall ceiling on 1080p MOT17 (median person ≈34 px): all YOLO sizes/models recall only 27–44% of visible pedestrians.

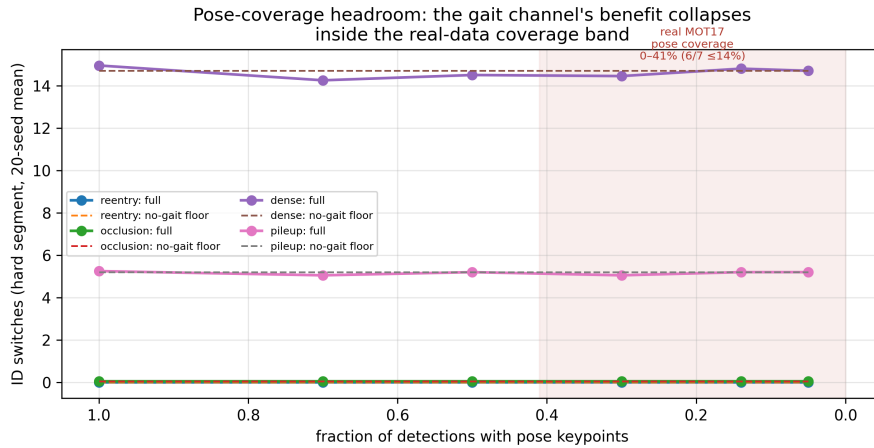


Figure S6: Pose-coverage headroom: the gait channel's switch benefit (full vs. −gait) as keypoint coverage falls. The benefit stays ≈0 at every coverage level on the synthetic hard segments — the decision layer resolves these events before the gait stage fires — which is the null result that motivates the real-data framing: the gait channel's measured contribution (44 switches, §2.5c) is bounded by the 0–41% real MOT17 coverage band (six of seven sequences at ≤14%), marked in gray.