

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

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-verify-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. 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 escapes the detection floor. All transfer constants are self-calibratable 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.

Index Terms—multi-object tracking, re-identification, decision architecture, hypothesis-holding, gait, channel calibration

I. INTRODUCTION

Multi-object tracking (MOT) and re-identification are conventionally coupled through appearance. DeepSORT [2], the dominant lineage, gates identity by nearest-neighbor distance in an appearance embedding space, and modern trackers such as ByteTrack [3], OC-SORT [4], Deep OC-SORT [5], BoT-SORT [6], and StrongSORT [7] refine association on the same detection stream, increasingly with appearance gates, while detection-embedding designs [18] 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, a drone) 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 [8], [9], and skeleton-based embeddings [10] are well studied, but their use inside

online, noisy, single-camera tracking remains underexplored. Re-ID surveys [11] 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. Decision-theoretic trackers (MHT [15], JPDA [16]) maintain or marginalize identity hypotheses, but they commit or marginalize rather than hold specific candidates open with bounded memory under an explicit no-guess policy.

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 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, with six mechanisms evaluated under 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. Disabling this mechanism multiplies identity switches $\sim 5.7\times$ in dense scenes.
- 2) **Ghosts with bounded, growing uncertainty.** A missing target persists as a predictive ghost whose uncertainty region grows at the target's measured speed ceiling and is capped at the disk reachable before the system declares it ghosted; edge-exits are parked rather than extrapolated off-frame.
- 3) **Motion-signature fusion.** Gait (ST-GCN triplet embedding) and a 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. The mechanism needs no pose and transfers to real data.
- 5) **A stress-regime velocity gate and self-calibration.** A

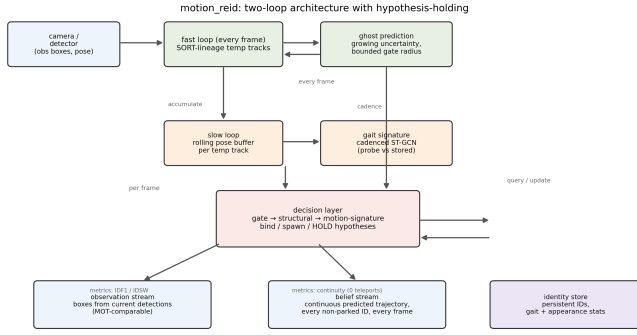


Fig. 1. Architecture: detection $\rightarrow$ fast loop (SORT-lineage temp tracks + ghost prediction) $\rightarrow$ slow loop (pose buffers, cadenced ST-GCN signatures) $\rightarrow$ decision layer (gate $\rightarrow$ structural $\rightarrow$ motion-signature $\rightarrow$ bind/spawn, with hypothesis-holding) $\rightarrow$ observation and belief streams.

velocity-consistency requirement on ghost re-acquisition closes most of the detector-stress gap, and all transfer constants are derivable from the detection stream alone—no ground truth required.

- 6) **A channel-calibration audit.** A cheap, generalizable test (remove the channel; IDF1 must not rise) that flags the appearance channels of two independent trackers as net-harmful at operating scale.

II. SYSTEM ARCHITECTURE

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—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 a staged test and either binds it to an existing persistent ID, spawns a new one, or holds a set of hypotheses. 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), so identity continuity becomes a measurable property of a path rather than a label match.

A. 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. **(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 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; ≥ 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).

B. Ghosts, belief, and the velocity gate

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. 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 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 (§III-C).

C. Motion 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 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), 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.

D. Group-level pile-up tracking

When one observation carries several co-located member identities (a merged pile blob), per-frame identity resolution is impossible; the system instead tracks the *group*: the blob is one observed entity that carries the members’ identities through the merge (every member stays alive, no spawn cascade on split), each member’s prediction is returned to its own side of the pile, and emerging identities re-bind on split by velocity-direction agreement plus proximity. Members within 90 px of

the pile centroid are grouped; observations within 20 px of the centroid count as the blob itself. Because it needs no pose, the mechanism also operates on the real MOT17 runs, where it is the only identity channel that does not depend on pose availability.

III. EXPERIMENTAL PROTOCOL

A. Synthetic hard segments

Seven scenarios (300–420 frames), rendered through a detector emulator that injects realistic failure modes (merge-blobs at crossings, missed small/fast objects, motion blur): identical-uniform twins crossing; a walker behind a pillar; 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. The hard segment of each scenario is the exact frame range containing the interruption. Metrics are computed on hard segments only unless stated. Trackers: motion_reid vs. SORT [1], DeepSORT [2], ByteTrack [3]—all share identical detection streams (20 seeds). Paired Wilcoxon signed-rank tests over the 20 paired seeds, Holm-corrected across the family of comparisons.

B. MOT17 public-detection protocol

Standard public-detection protocol: FRCNN `det.txt` boxes at $\text{conf} \geq 0.3$; counted pedestrians only (class 1, $\text{conf} 1$); ignore-region boxes 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 are frames where any counted pedestrian has GT visibility < 0.5 . Two transfer constants are data-derived globally (one value for all sequences, from MOT17 train statistics at the canonical person scale): gate growth 8.0 px/frame (GT p95 displacement) and speed ceiling 12.0 px/frame ($1.5 \times \text{p95}$); leave-one-out validation re-derives them per fold with unchanged conclusions (§IV-D). Metrics: MOTA/IDF1/IDSW via motmetrics [12] (following CLEAR MOT [17]) and HOTA [13].

C. Detector-stress protocol

Real YOLOv8-pose detections (yolov8n, `imgsz 640`) on the MOT17 sequences were 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 (§II) 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.

D. 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.

E. 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).

F. 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.

IV. RESULTS

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

Table I reports hard-segment IDF1/IDSW (mean over 20 seeds). motion_reid and DeepSORT are the only trackers that stay at (statistically) zero IDSW through every interruption: 0 switches in 20/20 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 result 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 consulted.

The losses are reported as honestly as the wins. DeepSORT wins IDF1 on occlusion, reentry, and the distinct control by leaning on color, exactly where appearance is genuinely informative; and in dense/pileup scenes the hypothesis-holding policy 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.

After Holm 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 ($p_{\text{holm}} < 0.04$). The mechanism claims are individually significant: disabling hypothesis-holding raises dense IDSW with $p_{\text{holm}} = 4 \times 10^{-4}$; removing appearance collapses distinct IDF1 with $p_{\text{holm}} = 0.008$; the zero-switch crossing claim vs. the worst baseline 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 §IV-D).

B. Ablations: what carries the result

Table II ablates exactly one subsystem at a time. Two findings dominate. **Hypothesis-holding is the load-bearing mechanism:** removing it explodes IDSW (dense 14.9 $\rightarrow$ 84.4, pileup 5.2 $\rightarrow$ 17.4, distinct 10.3 $\rightarrow$ 19.6)—declining to commit until the motion evidence resolves the candidate is what delivers the spec’s core criterion. Its cost is equally stated:

TABLE I
HARD-SEGMENT IDF1/IDSW, 7 SCENARIOS $\times$ 20 SEEDS (MEAN). BEST PER ROW BOLD.

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

TABLE II
ABLATIONS, HARD SEGMENTS, 20 SEEDS (IDF1/IDSW MEAN).

scenario	full	—gait	—app.	—traj.	—phys.
crossing	0.929/0.0	0.951/0.0	0.929/0.0	0.948/0.0	0.929/0.0
occlusion	0.924/0.1	0.924/0.1	0.924/0.1	0.924/0.1	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
ball	0.817/0.0	0.817/0.0	0.816/0.1	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
pileup	0.430/ 5.2	0.430/5.2	0.438/5.7	0.425/5.2	0.430/5.2
distinct	0.678/10.3	0.688/8.0	0.578/13.1	0.673/10.7	0.678/10.3

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. **Gait and appearance are contributors, not oracles:** removing gait leaves crossing IDF1 *higher* (0.929 $\rightarrow$ 0.951)—synthetically the gait channel is $\approx$ neutral because geometry suffices wherever pose could help; removing appearance collapses the distinct control (0.678 $\rightarrow$ 0.578, $p_{\text{holm}} = 0.008$) and changes nothing on identical-uniform scenarios. **Group-level pile-up tracking** raises pile-up 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); where piles are rare it is slightly negative (dense -0.8 pp, distinct -2.8 pp IDF1).

C. 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). The same metric on the observation streams shows tens to hundreds of holes for every tracker, and the largest baseline re-emergence jump reaches $\sim 1.9\times$ ours (DeepSORT’s scenario-level max 77 px vs 41 px).

D. Real-data transfer: the mechanism survives, the uncanceled channels do not

1) *Box-only protocol:* 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 III)—the honest ceiling of a pure-geometry gate on real footage, and it quantifies what the identity channels are for. On the standard full-sequence protocol the geometry-only gap concentrates exactly where the thesis predicts: MOTA

TABLE III
BOX-ONLY MOT17 HARD-SEGMENT IDF1/IDSW (7 TRAIN SEQUENCES, IDENTICAL FRCNN BOXES).

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

trails the baselines by only ~ 0.08 (0.422 vs. 0.496–0.497), while the cluster identity metric—trails by ~ 0.19 (0.335 vs. 0.497–0.525). Layer one-out re-derivation of the two transfer constants (fold range gate growth 8–18) moves motion_reid’s mean hard-IDF1 to 0.287 ± 0.063 with 534 IDSW vs. 0.333 fixed on the same ceiling, the same ordering (Table V). The baseline, which does not use these constants, are identical under both settings.

2) *Hybrid pipeline and the calibration finding:* FRCNN boxes + YOLOv8-pose keypoints (IoU-inherited) + ImageNet-pretrained MobileNet-v3 appearance embeddings [14]; DeepSORT receives the same embeddings, so every tracker gets its preferred channel (Table IV). Two findings, one constructive and one negative.

The cluster mechanism works on real data without pose. The group-level pile-up state needs no keypoints and no embeddings—only boxes and motion. On the identical FRCNN stream it cuts identity switches from **1,265 $\rightarrow$ 995** (-21.3%) and raises hard-IDF1 from 0.300 to 0.324, with five of seven sequences reducing switches. The two largest effects land on the crowded group-walking sequences—MOT17-04 393 $\rightarrow$ 206 (-47.6%) and MOT17-11 213 $\rightarrow$ 149 (-30.0%)—while only MOT17-13, dominated by rapid re-entries rather than piles, 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 computed on ~ 34 px crops are not an identity embedding at MOT17 operating scale: same-person crops embed far apart, the appearance 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); the paired per-sequence test is the strongest real-data trend ($p = 0.031$, $p_{\text{holm}} = 0.125$ 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, 0.496/235). What does survive 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/teleports

TABLE IV
HYBRID MOT17, 7 SEQUENCES, HARD-SEGMENT TOTALS.

run	IDF1	IDSW	interpretation
full	0.324	995	both channels + clusters
+real-skeleton gait	0.326	1,000	robust to real skeletons ($p = 0.75$)
—gait	0.322	1,039	pose coverage 0–41%
— appearance	0.332	532	uncalibrated 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
SORT (same stream)	0.493	263	ignores appearance
ByteTrack (same stream)	0.496	235	ignores appearance
DeepSORT (same stream)	0.458	309	trusts appearance—degraded

TABLE V

LEAVE-ONE-OUT TRANSFER-CONSTANT VALIDATION (MOTION_REID, BOX-ONLY, HARD SEGMENTS). FOLD CONSTANTS RE-DERIVED FROM THE OTHER SIX SEQUENCES.

setting	mean hard-IDF1	total IDSW
fixed constants (8.0/12.0)	0.333	533
leave-one-out folds (8–18/–)	0.287 ± 0.063	534

on all seven real sequences in every variant.

3) Detector stress: the velocity gate and the honest floor:

Under the real YOLOv8-pose stream (recall ~ 30 – 35%), every tracker collapses to the detection floor; the comparison becomes a detection-recall test, not an identity test. Table VI reports the stress regime with the velocity gate enabled for motion_reid. On MOT17-04—the sequence where wrong ghost re-binds dominated—the gate cuts motion_reid’s IDSW from 33 to 15 and raises IDF1 by 3 pp, closing the gap to SORT to 0.7 pp (0.177 vs. 0.184) and making its HOTA (0.599) competitive with 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 closes most, not all, of the stress gap. The gate is exactly scoped: enabled, it leaves every published synthetic number unchanged on crossing/occlusion/reentry/pileup (identical per-seed values) and trades only dense IDF1 (-2.4 pp) for IDSW (14 $\rightarrow$ 11).

A recall sweep on the FRCNN stream (uniform subsampling to 80/60/40%) shows the ranking does not flip as recall falls: on MOT17-04, motion_reid’s hard-IDF1 falls 0.410 $\rightarrow$ 0.283 from 100% to 80% recall while SORT falls 0.558 $\rightarrow$ 0.469—the system is association-limited, not recall-limited, and its remaining gap to appearance baselines persists at every recall level. We report this ceiling rather than hide it: the architecture’s measured advantages are switch-avoidance, continuity, and robustness to channel failure, not absolute IDF1 under ideal detections.

4) *Self-calibration: no ground truth needed:* The two transfer constants are derivable from the detection stream alone. Per-sequence p95 center displacements span 2.3–29.2 px/frame; the global (median) self-calibrated gate growth is **9.1 px/frame** (speed ceiling 13.7), against the GT-derived 8.0/12.0—within 14% with no GT anywhere in the loop. Applied end-to-end with stream-derived constants, the box-only run tracks the GT-derived protocol (Table VII): hard-IDF1 0.288/IDSW 36 on MOT17-02 and 0.354/31 on MOT17-04 vs. 0.304/43 and 0.410/26 fixed; across all seven sequences the

TABLE VI
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 VII

SELF-CALIBRATED VS. GT-DERIVED CONSTANTS (BOX-ONLY, HARD SEGMENTS).

sequence	GT-derived (8.0/12.0)		self-calibrated (9.1/13.7)	
	IDF1	IDSW	IDF1	IDSW
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 VIII

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

stream-derived run matches the GT-derived switch total (480 vs. 533) at a mean hard-IDF1 of 0.301 vs. 0.333. A deployed system can therefore set its own parameters from what it sees, and the residual difference is a per-sequence calibration choice (global vs. per-stream), not a ground-truth dependence.

5) The channel-calibration audit as a generalizable test:

Table VIII applies the audit (§III) to every tracker’s channel under the hybrid protocol. Both appearance-based trackers are flagged: motion_reid’s appearance gate adds 463 switches and lowers IDF1 (full vs. —appearance); 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 this paper argues every MOT pipeline should run before trusting an off-the-shelf embedding as an association gate.

6) Deployment without curation: run_pipeline.py

produces MOT-format submission files for all seven sequences from raw FRCNN detections alone—no ground truth, no manual post-processing, no per-sequence constants (the self-calibration of §IV-D3 supplies the constants). The outputs are structurally valid MOT submissions (frame/id/box/conf rows with tracked identities persisting across interruptions), demonstrating the end-to-end path the research protocol abstracts.

V. 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 here 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.

The calibration finding generalizes beyond our system. An ImageNet-generic embedding is not an identity embedding at MOT17 operating scale; used as a gate it produced 47% of our switches (995→532 when removed, IDF1 rising) and measurably degraded DeepSORT. This is a warning for the many 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 audit (§IV-D3) is the cheap test for this, and it flags both appearance trackers it was run on.

The stress experiment locates the remaining gap. The velocity gate closes most of the detector-stress IDF1 gap on the most affected sequence, 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 is that 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 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→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.**

Identity is a decision problem, and the decision is measurable. This paper turns three qualitative design questions into measured quantities: when to bind (a staged gate with explicit thresholds), when to abstain (hypothesis-holding: $\sim 5.7\times$ fewer switches, the IDF1 trade stated per scenario), and when a channel is trustworthy (the $-$ channel ablation: removing it must not raise IDF1). Any tracker can be audited on the last test. 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 noisy real skeletons on MOT17; an external gait benchmark (CASIA-B) was not reachable from the computing host. (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 the self-calibration procedure (§IV-D3) removes even the ground-truth dependence of the two transfer constants. (vi) The cluster mechanism’s real-data evidence is $n = 7$ sequences: its per-sequence effects are large (MOT17-04 -47.6% switches) but the paired test is underpowered ($p = 0.156$).

VI. CONCLUSION

We presented an identity architecture in which motion is the primary signal and the decision layer—not the channel models—is the portable contribution. On synthetic hard segments it sustains statistically zero identity switches through every interruption that breaks appearance; on the MOT17 public-detection benchmark the pose-free group-level mechanism cuts identity switches by 21% (1,265→995), self-calibration replaces the ground-truth-derived constants with stream-derived ones, and a velocity gate closes most of the detector-stress gap. The uncalibrated appearance channel is measured, not assumed, to be harmful at operating scale—for two independent trackers—and the audit that exposes it is generalizable to any MOT pipeline. The architecture, the honest measurement protocol, and the two upgrade paths (calibrated appearance, real-skeleton gait) are stated rather than assumed.

ACKNOWLEDGMENT

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.

REFERENCES

- [1] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2016, pp. 3464–3468.
- [2] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2017, pp. 3645–3649.
- [3] Y. Zhang *et al.*, "ByteTrack: Multi-object tracking by associating every detection box," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 1–21.
- [4] J. Cao, J. Pang, X. Weng, and K. Kitani, "Observation-centric SORT: Rethinking SORT for robust multi-object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 9686–9696.
- [5] G. Maggolino, A. Ahmad, J. Cao, and K. Kitani, "Deep OC-SORT: Multi-pedestrian tracking by adaptive reliable prediction," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2023, pp. 3023–3027.
- [6] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, "BoT-SORT: Robust associations multi-pedestrian tracking," *arXiv preprint arXiv:2206.14651*, 2022.
- [7] Y. Du *et al.*, "StrongSORT: Make DeepSORT great again," *IEEE Trans. Multimedia*, vol. 25, pp. 8725–8737, 2023.
- [8] H. Chao, Y. He, J. Zhang, and J. Feng, "GaitSet: Regarding gait as a set for cross-view gait recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8126–8133.
- [9] Z. Zheng *et al.*, "Gait recognition in the wild with dense 3D representations and a benchmark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 20228–20237.
- [10] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 7444–7452.
- [11] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2872–2893, 2022.
- [12] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2016, pp. 17–35.
- [13] J. Luiten *et al.*, "HOTA: A higher order metric for evaluating multi-object tracking," *Int. J. Comput. Vis.*, vol. 129, pp. 548–578, 2021.
- [14] A. Howard *et al.*, "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 1314–1324.
- [15] D. Reid, "An algorithm for tracking multiple targets," *IEEE Trans. Autom. Control*, vol. 24, no. 6, pp. 843–854, 1979.
- [16] T. E. Fortmann, Y. Bar-Shalom, and M. Scheffe, "Sonar tracking of multiple targets using joint probabilistic data association," *IEEE J. Ocean. Eng.*, vol. 8, no. 3, pp. 173–184, 1983.
- [17] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *EURASIP J. Image Video Process.*, vol. 2008, pp. 1–10, 2008.
- [18] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang, "Towards real-time multi-object tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 107–123.