ACTIVATE SPARSE, PREDICT CONTINUOUSLY: BRAIN-INSPIRED TERNARY SPARSE ACTIVATION WITH PREDICTIVE CODING FOR MULTI-OBJECT TRACKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-object tracking (MOT) must detect objects and preserve their identities in continuous video streams under tight computational budgets. Spiking neural networks (SNNs) promise low-power computation, but their sparsity lives along the temporal axis: they encode information by accumulating events over timesteps, which mismatches the instantaneous spatial discrimination and continuous state estimation that online MOT demands, and occlusion-induced event loss destroys the features needed to re-identify a target. To address these issues, this paper proposes an efficient online tracking framework inspired by the sparse coding and predictive cognition mechanisms of biological vision, termed Activate Sparse, Predict Continuously (ASPC). First, we propose a ternary sparse activation network that gates neural responses dynamically and encodes features with ternary sparse activations $(-1/0/+1)$, shifting sparse computation from the temporal domain to a ternary sparse activation-based feature representation space; this mechanism preserves the directional discriminative information of single-frame features while enabling adaptive feature selection. Furthermore, we design a prediction-driven cognitive association module that formulates the tracking process as dynamic inference based on historical state priors, thereby maintaining identity continuity through internal states when visual information is insufficient. Experiments on MOT17, DanceTrack, and SportsMOT show that ASPC matches or surpasses the strongest full-precision trackers in HOTA and AssA while using a single forward pass: on MOT17 it attains HOTA 59.3 and AssA 58.2 at 25.7 mJ per frame, against 1105 mJ for the strongest full-precision baseline, and it exceeds an identically configured spiking counterpart by 29.3 HOTA points on MOT17 and 22.5 points on DanceTrack. Its activation sparsity adapts to scene density (48–59% zero activations), with per-frame inference energy of only 25.7 mJ, 11 mJ, and 19 mJ on the three datasets.

1 INTRODUCTION

extbfOnline tracking and its cost. Multi-object tracking (MOT) requires detecting objects in every frame of a video stream while preserving each object's identity across frames, under appearance changes, motion, occlusion, and crowding. Online MOT must decide instantaneously as each frame arrives. State-of-the-art trackers are built on the tracking-by-detection paradigm (Bergmann et al., 2019; Wojke et al., 2017): a detector proposes candidates, specialized models extract appearance and motion cues, and an association module links detections to existing tracks. While modular and accurate, these pipelines are dominated by heavyweight deep networks whose compute and energy costs grow rapidly with resolution and frame rate, making deployment on resource-constrained devices – embedded cameras, drones, battery-powered sensors – a major open problem.

extbfTwo sparse paradigms, two mismatches. Sparsity is the natural lever for reducing this cost, yet the two main sparse paradigms are each misaligned with online MOT in a specific way. Spiking neural networks (SNNs) communicate only when neurons fire, and event-driven hardware can exploit this to cut both computation and memory traffic (Davies et al., 2018; Roy et al., 2019). However, SNN sparsity lives along the temporal axis: accuracy is recovered by accumulating spikes over multiple timesteps, trading latency for precision (Deng & Gu, 2021; Bu et al., 2022) – precisely the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

opposite of what online tracking needs, which is instantaneous spatial discrimination within a single frame. More fundamentally, SNN identity information resides in membrane potentials that persist only while events keep arriving; when an object is occluded, its events stop, the potential leaks away, erasing the very features needed to re-identify the object upon reappearance. Binary networks remove multiplications (Courbariaux et al., 2016; Rastegari et al., 2016) but offer no exploitable sparsity: unipolar codes lose sign information, and bipolar codes contain no zeros. What online MOT needs is therefore not sparsity in time but sparsity in the feature space: per-frame features that are intrinsically sparse, signed, and produced in a single forward pass.

extbfLessons from biological vision. Biological vision resolves exactly this combination of constraints. Under severe metabolic budgets, primate cortex produces stable, discriminative perception from sparse, noisy, interrupted input via two complementary principles. First, sparse signed coding: only a small fraction of neurons respond to any stimulus, with responses carrying sign, so directional contrast is preserved at minimal firing cost (Olshausen & Field, 1996). Second, predictive processing, formalized by the free-energy principle: the brain maintains an internal generative model and continuously predicts incoming sensory data, recasting perception as inference in which posterior beliefs balance evidence fit against the prior supplied by the internal model (Friston, 2010; Rao & Ballard, 1999). When evidence is absent, as during occlusion, the prediction dominates – which is why perceived objects persist through gaps in observation. These two principles map directly onto the two bottlenecks of efficient online MOT: sparse signed coding addresses the cost of feature representation, and predictive inference addresses identity stability under missing evidence.

We introduce ASPC (Activate Sparse, Predict Continuously), an online tracking framework with two components. (i) A ternary sparse activation network gates each neuron's response into -1, 0, +1: silent units contribute no computation or memory access, the signed active values retain the directional discriminability that unipolar binary codes discard, and the gate performs adaptive feature selection. Since this sparsity is intrinsic to the spatial representation, ASPC gains the energy benefits of event-driven computation without multi-timestep latency, and its sparsity adapts to scene density. (ii) A prediction-driven cognitive association module formulates track–detection association as dynamic inference over persistent internal states: each track carries a state propagated by a dynamics prior and updated by current-frame cues; when visual evidence is insufficient, the module falls back on the internal state, emitting calibrated soft assignments rather than committing to a hard, potentially wrong match, so identity is carried through gaps in observation.

On MOT17, DanceTrack, and SportsMOT, ASPC reaches HOTA 59.3, 61.2, and 75.5 in a single forward pass, exceeding the identically configured SNN counterpart by 29.3, 22.5, and 19.8 HOTA points while cutting its per-frame energy by up to $24\times$, with sparsity adapting to scene density (48–59% zero activations). Our contributions are:

- We identify a fundamental mismatch between the temporal sparsity of SNNs and the instantaneous, occlusion-robust requirements of online MOT, and relocate sparsity from the time domain to the feature representation space.
- We design a ternary sparse activation network with dynamic gating that encodes single-frame features with signed ternary codes, preserving directional discriminability while enabling scene-adaptive feature selection.
- We formulate track association as prediction-driven dynamic inference over persistent internal states, providing a principled mechanism for identity continuity under missing visual evidence and yielding calibrated soft association decisions.
- We show that ASPC substantially improves association quality over its SNN counterpart at a fraction of the energy, establishing a new operating point on the accuracy–efficiency frontier for online tracking.

2 Related Work

Tracking-by-detection with heuristic association. The dominant online paradigm is tracking-by-detection (TbD), rooted in SORT (Bewley et al., 2016) and DeepSORT (Wojke et al., 2017). Successive work refines three families of hand-crafted rules: tracklet representation (Wojke et al., 2017; Zhang et al., 2021), motion–appearance feature fusion (Zhang et al., 2022b; Aharon et al.,

2022; Seidenschwarz et al., 2023; Maggiolino et al., 2023; Yang et al., 2024), and cascaded multi-stage matching (Zhang et al., 2022b; Cao et al., 2023; Aharon et al., 2022; Maggiolino et al., 2023; Qin et al., 2023; Lv et al., 2024). These heuristics encode strong inductive biases, but their reliability fluctuates with context, and the oracle study of CamelTrack (Somers et al., 2025) shows that replacing the association heuristic with an optimal oracle improves HOTA by 15.5% and 8.3% on DanceTrack and SportsMOT: heuristic association, not cue quality, is the binding constraint.

Learned association. End-to-end methods re-detect past objects with track queries (Meinhardt et al., 2022; Zeng et al., 2022; Zhang et al., 2023; Gao & Wang, 2023; Gao et al., 2024), learning detection and association jointly at the cost of antagonistic objectives, no support for off-the-shelf models, and multi-day training. Within TbD, TransMOT and TADN attach learned components to heuristic stages (Chu et al., 2023; Psalta et al., 2024), and CamelTrack (Somers et al., 2025) is the first fully learned, heuristic-free association module. ASPC builds on CamelTrack and asks whether the association module itself can be made sparse and predictive, so that learned-association accuracy is obtained at the energy arithmetic of event-driven computation.

Sparse and spiking networks. Binary networks remove multiplications (Hubara et al., 2016; Liu et al., 2018; 2020; Zhang et al., 2022a; Tu et al., 2022) but bipolar codes contain no zeros and unipolar codes lose sign (Wang et al., 2020; Sakr et al., 2019); adder networks remain dense (Chen et al., 2020). SNNs achieve genuine event-driven sparsity (Davies et al., 2018; Roy et al., 2019; Kim et al., 2020; Neftci et al., 2019), yet their accuracy is recovered by accumulating spikes over timesteps (Deng & Gu, 2021; Bu et al., 2022; Zheng et al., 2021; Chowdhury et al., 2021; Deng et al., 2022; Rathi et al., 2020)—the opposite of per-frame discrimination. Recent work improves SNN training with high-order information bottleneck objectives (Yang & Chen, 2025) and extends spiking models to event-based optical flow (Yang et al., 2025) and to the efficient training of neuromorphic electronics (Yang et al., 2026). Most closely related, binary activation networks gate activations at the Hoyer extremum of their magnitude map (Datta et al., 2024; Kurtz et al., 2020); our ternary network extends this line by restoring signed responses and applying the mechanism to a tracker's feature representations.

Predictive coding and free energy. Predictive coding frames cortical computation as top-down prediction of bottom-up evidence (Rao & Ballard, 1999), and the free-energy principle generalizes this to variational inference under an internal generative model (Friston, 2010; Friston et al., 2015; Nuijten et al., 2026). Kalman filtering in trackers (Kalman, 1960) shares this generative spirit but is used only as a cue-producing heuristic. ASPC instead makes prediction-driven inference the association decision rule itself: the predicted state is the prior, the current embedding is evidence, and matching minimizes their free energy.

3 METHOD

We keep CamelTrack's online tracking-by-detection pipeline and replace its association core with two coupled mechanisms (Fig. 1): (i) a ternary sparse activation network (TSAN) that turns the dense floating-point activations of the association module into gated ternary codes, and (ii) a prediction-driven cognitive association module (PAC) that performs track–detection matching as variational free energy minimization over internal track states. We first fix notation, then detail each mechanism, then the training objective and energy accounting.

3.1 PROBLEM DEFINITION

Definition 1Online multi-object tracking. *Given a video stream $V = \{I_1, I_2, \dots\}$ and an off-the-shelf detector producing detections $\mathcal{D}^t = \{d_1^t, \dots, d_{N_t}^t\}$ at each frame t, online multi-object tracking seeks, at every frame t and using only frames up to t, an assignment $\pi^t : \{1, \dots, M^t\} \to \{1, \dots, N_t\} \cup \{\varnothing\}$ between the M^t active tracklets $\mathcal{T}^t$ and the current detections, such that each detection is matched to at most one tracklet and vice versa, and detections matched to the same tracklet across frames belong to the same object. The quality of π^t is measured by HOTA/AssA/IDF1, and its cost by per-frame inference energy.*

Remark.*Unlike offline tracking, π^t is irrevocable: a wrong match corrupts the tracklet's feature bank and propagates to subsequent decisions, making calibrated, near-boundary decisions the central algorithmic requirement, addressed by the module of Sec. 3.4.*

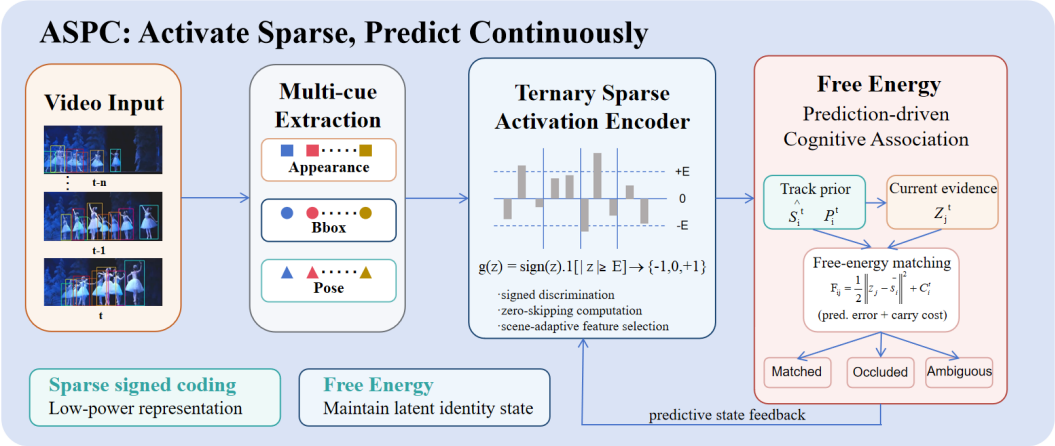

Figure 1: Overview of the proposed ASPC framework. Ternary-gated encoders produce sparse signed embeddings, the prediction-driven match gates each pairing at the learned threshold, and matched tracks fold evidence back into the feature bank for the next prediction step.

3.2 Overview and Notation

At frame t an off-the-shelf detector produces detections $\mathcal{D}^t = \{d_j^t\}$; for each detection, $K{+}1$ cues $c_{k,j}^t$ are extracted (bounding box plus confidence for $k{=}0$; appearance and pose cues for $k{=}1, 2$), so $d_j^t = \{c_{k,j}^t\}_{k=0}^K$. The association step must match M existing tracklets $\mathcal{T} = \{\tau_i\}$ to the N current detections. Each tracklet is a sequence of detections with a common identity and keeps a feature bank of its W most recent frames, $h_i^{t-W:t-1}$, which the tracker maintains across frames; this bank is the substrate for the track's internal state (Sec. 3.3). ASPC processes the active set $\mathcal{A} = \mathcal{T} \cup \mathcal{D}^t$ in two encoding stages, exactly as in CamelTrack (Somers et al., 2025): each per-cue Temporal Encoder TE_k aggregates the sequence $\{c_{k,i}^{t_s}, \ldots, c_{k,i}^{t_e}\}$ of cue k for object i into a temporal token $y_{k,i}$ (with sinusoidal age encodings and a learned [CLS] readout), and GAFFE sums linear projections of the cue tokens, $\hat{y}_i = \sum_{k=0}^K \text{Linear}_k(y_{k,i})$, then refines the multi-modal tokens through inter-object self-attention (Vaswani et al., 2017) into unified embeddings $\{z_i\}_{i\in\mathcal{A}} \in \mathbb{R}^d$. Matched objects should be close in this space; a cost matrix $C_{ij} = \|z_i - z_j\|_2$ drives bipartite matching with the Hungarian algorithm (Kuhn, 1955). The two design changes we introduce—ternary gating and free-energy-based decisions—are described next.

3.3 Ternary Sparse Activation Network

Motivated by sparse signed coding in biological vision, where only a small fraction of neurons respond to any stimulus and their responses carry sign (Olshausen & Field, 1996), all dense layers of TE and GAFFE (cue projections, attention feed-forwards, readouts) share the same gated activation. For a layer l with pre-activation tensor $u_l = W_l\, o_{l-1}$, we normalize it by a trainable per-layer scale v_l (the *type threshold*) to obtain $z_l = u_l / v_l$, and compute the Hoyer extremum of its magnitude map,

$$E_l \;=\; \frac{\|m_l\|_2^2}{\|m_l\|_1}, \qquad m_l = |z_l|, \tag{1}$$

which is the non-negative tensor value at which the Hoyer sparseness measure $\mathcal{H}(m) = \big(\|m\|_1/\|m\|_2\big)^2$ attains its minimum, i.e., the point of maximal marginal response per unit activation mass (Datta et al., 2024; Hoyer, 2004). The ternary gate then maps each response to a signed trit:

$$o_l = g(z_l) = \text{sign}(z_l) \cdot \mathbf{1}\big[\,|z_l| \geq E_l\,\big] \quad \in \{-1, 0, +1\}, \tag{2}$$

so that E_l is the magnitude boundary between silence and firing, chosen from the data distribution itself rather than fixed *a priori*. Three properties follow. (i) *Exploitable sparse computation:* where

$o_l = 0$, the corresponding weight columns of layer $l{+}1$ are skipped entirely—no multiply, no accumulate, no weight fetch; our measured per-layer zero ratios are 48–59% (Tables 1–3). (ii) *Signed codes:* active units carry direction, so single-frame features retain the contrast structure that unipolar $\{0, 1\}$ codes discard (Wang et al., 2020; Sakr et al., 2019) and that the downstream association needs to compare tracklets and detections in a metric space. (iii) *Data-adaptive selection:* with $E_l = \|m_l\|_2^2/\|m_l\|_1$ and $\|z_l\|$ normalized by the trained scale v_l, the gate fires exactly the subset of units whose magnitudes lie above the marginal-mass boundary of the current layer's response distribution; features uninformative for the current scene are silenced because their mass concentrates below the boundary. Because each dataset learns its own v_l and E_l, the resulting sparsity adapts to scene statistics with no per-scene hyper-parameter tuning.

Training through the gate uses the standard straight-through estimator in surrogate form (Datta et al., 2024): the gradient of Eq. 2 is approximated by a box of width E_l centered on the boundaries,

$$\frac{\partial o_l}{\partial z_l} \;\approx\; s \cdot \mathbf{1}\Big[\,\big|\,|z_l| - E_l\big| < E_l\,\Big], \tag{3}$$

with s a damping scale; units gated off receive no gradient, so weight updates only flow through active units—the backward pass inherits the forward sparsity. Following BANN (Datta et al., 2024; Kurtz et al., 2020), we regularize the magnitude map of each gated layer with the scale-invariant Hoyer regularizer (Hoyer, 2004; Yang et al., 2020)

$$\mathcal{L}_H \;=\; \sum_l \mathcal{H}(m_l) \;=\; \sum_l \Big(\tfrac{\|m_l\|_1}{\|m_l\|_2}\Big)^2, \tag{4}$$

weighted by λ_H in the total loss (Eq. 10). Its gradient pushes magnitudes away from E_l—the minimum of $\mathcal{H}$ —which displaces the activation population away from the decision boundary and makes the gate robust to the noise that would otherwise flip binary decisions near threshold (Datta et al., 2024). At inference we use the exponential moving average of each layer's E_l collected during training, exactly as batch-normalization statistics. The first layer receives full-precision input and is not gated, as in all sparse-multiplier realizations (Datta et al., 2024).

3.4 Prediction-driven cognitive association

The second mechanism replaces the association decision with prediction-driven inference. Building on the free-energy principle of cortical computation (Friston, 2010; Rao & Ballard, 1999), we recast matching as variational free energy minimization, giving identity continuity a principled substrate. We model each tracklet τ_i as carrying a hidden state $s_i^t \in \mathbb{R}^d$—its persistent identity representation—and factor the tracking problem into the generative model

$$p(s_i^t \mid s_i^{t-1}) \;=\; \mathcal{N}\big(\bar{s}_i^t,\; P_i^t\big), \qquad p(z_j^t \mid s_i^t) \;=\; \mathcal{N}\big(s_i^t,\; \Sigma_j^t\big), \tag{5}$$

where the *dynamics prior* $p(s_i^t|s_i^{t-1})$ is supplied by the track's feature bank: the temporal encoder ran over $h_i^{t-W:t-1}$ is the predicted state $\bar{s}_i^t = \mathrm{TE}(h_i^{t-W:t-1})$, and its predictive covariance P_i^t grows with the age of the freshest evidence in the bank (a track that has been missing for long predicts diffusely). The current-frame embedding z_j^t produced by GAFFE is the *likelihood evidence*: if detection j continues identity i, its embedding concentrates around s_i^t with measurement covariance Σ_j^t. Under the variational posterior $q(s_i^t)$ over the track's state, the free energy of binding detection j to tracklet i decomposes into a prediction-error (accuracy) term and a complexity (prior-fit) term:

$$F_{ij}^t \;=\; \underbrace{\tfrac{1}{2}\big\|z_j^t - \bar{s}_i^t\big\|_{\Sigma^{-1}}^2}_{\text{prediction error}} + \underbrace{C_i^t}_{\text{state complexity (carry cost)}}, \tag{6}$$

where $\|x\|_{\Sigma^{-1}}^2 = x^\top \Sigma^{-1} x$ and the carry cost C_i^t encodes how far the posterior must depart from the dynamics prior to explain the current frame—it is small for a confident, freshly updated track and large for a stale one. Association is then a free-energy *comparison* against the null hypothesis that j

opens a new identity with fixed spawn cost γ: assign $j \to i$ when $F_{ij}^t < \gamma$. With isotropic evidence covariance $\Sigma_j^t = \sigma^2 I$ this reduces to a similarity threshold,

$$\left\| z_j^t - \bar{s}_i^t \right\|_2^2 \; < \; \theta_i^t, \qquad \theta_i^t = 2\sigma^2 \big(\gamma - C_i^t\big), \tag{7}$$

i.e., the matching gate is the point where evidence fit and prior cost balance. This is exactly the threshold our tracker learns and calibrates from data—the model locates its own decision boundary instead of inheriting one from heuristics (the calibrated values differ markedly across datasets; Sec. 4.2). Eq. 7 also recovers the age-gating heuristics of SORT-style trackers (Wojke et al., 2017; Zhang et al., 2022b) as a special case: a long-missed track has large C_i^t and therefore demands stronger evidence to be re-associated, which is Bayes-sensible rather than hand-tuned.

The free-energy view prescribes the full recursion. *State update:* the matched posterior is the precision-weighted fusion of prediction and evidence, in the algebraic form of a Kalman gain,

$$s_i^t \; = \; \Sigma_{\text{post}} \left(\Sigma^{-1} z_j^t \; + \; P_i^{-1} \bar{s}_i^t \right), \qquad \Sigma_{\text{post}}^{-1} = \Sigma^{-1} + P_i^{-1}, \tag{8}$$

so the next frame's prior inherits a corrected state that fuses all cues through the temporal encoder; GAFFE's inter-object self-attention, which normalizes each embedding against the competing active objects, plays the role of the local normalizer of the assignment posterior. *Missing evidence:* when a tracklet has no plausible detection—occlusion, exit, detector miss—the likelihood is non-informative ($\Sigma \to \infty$), equation 8 collapses to $s_i^t = \bar{s}_i^t$, and identity is propagated by the prior alone; the features needed to re-identify the object upon reappearance are never discarded, unlike a leaking membrane potential. *Calibrated soft assignment:* rather than committing to the Hungarian solution unconditionally, we expose the assignment posterior.

$$p(m_{ij} \mid \mathcal{A}^t) \; = \; \text{softmax}_j \big(- F_{ij}^t / \tau \big), \tag{9}$$

with temperature τ; the tracker commits to a hard match only when the margin between the best and second-best hypotheses exceeds a calibrated gap, and otherwise defers to the propagated state. Soft near-boundary decisions prevent the error cascades that hard wrong matches induce by corrupting feature banks (Somers et al., 2025).

3.5 Training objective and energy accounting

Objective. We train with CamelTrack's association-centric training (ACT) (Somers et al., 2025): an image-free training set is pre-generated by running the off-the-shelf detector, ReID, and pose models once over the training sequences, assigning each detection the identity of its IoU-closest ground truth; batches then sample cross-video association scenarios of P tracklet–detection pairs, enriched by identity-swap, detection-dropout, and cue-dropout augmentations. On each scenario we minimize.

$$\mathcal{L}_{\text{total}} \; = \; \mathcal{L}_{\text{NCE}} \; + \; \lambda_H \, \mathcal{L}_H, \tag{10}$$

where $\mathcal{L}_{\text{NCE}}$ is the InfoNCE loss (van den Oord et al., 2018) over paired tracklet–detection embeddings, which admits a variational reading: its negation is a lower bound on the mutual information between track state and detection evidence (Poole et al., 2019), so minimizing the contrastive loss *is* minimizing a free energy over the identity posterior—the same principle the association module instantiates at inference time. The Hoyer term $\mathcal{L}_H$ (Eq. 4) shapes the gated layers' pre-activation distributions. With $\lambda_H{=}10^{-9}$ (swept in Sec. 4.3) its numerical contribution to the objective is negligible ($\approx 5{\times}10^{-7}$—the final MOT17 training loss is 0.181 while the weighted Hoyer contribution is $\lambda_H \cdot 484 \approx 4.8{\times}10^{-7}$); its role is geometric—pushing magnitudes away from the Hoyer extremum— not to dominate the empirical loss, matching the mechanism by which Hoyer regularization separates binary thresholds in BANN (Datta et al., 2024). All models are trained from scratch (random initialization) for 30 epochs with AdamW, one GPU, no pre-training: MOT17 converges in 0.66 h, DanceTrack in 13.4 h, SportsMOT in 18.4 h—an order of magnitude less compute than E2E trackers (Zeng et al., 2022; Gao & Wang, 2023). The matching threshold of Eq. 7 is calibrated

on the validation split to maximize assignment accuracy, and its per-epoch evolution is monitored (Appendix C).

Energy accounting. Per-frame energy is measured on the target hardware during the final validation runs, with the measured idle baseline subtracted; full measurement details are given in Appendix B.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate on MOT17 (Milan et al., 2016), DanceTrack (Sun et al., 2022), and SportsMOT (Cui et al., 2023). Following the official splits: MOT17 (SDP variant) train/val splits with a held-out sequence as validation (leave-one-out index 5, i.e., MOT17-11-SDP; 9436 GT pedestrian boxes); DanceTrack's official validation set of 25 sequences (225148 GT boxes); SportsMOT's official validation set of 45 sequences (295573 GT boxes). States (detection boxes plus ReID and pose cues) are pre-generated for the association-centric training. To keep one uniform and reproducible recipe, all three datasets use the yolo11m-pose detector (Jocher & Qiu, 2025) as the detection source (pose cues from RTMPose-style estimators (Jiang et al., 2023); with up-sampled 1920-px inference at confidence 0.05 for SportsMOT's distant, small targets). This differs from CamelTrack's curated YOLOX detections (Somers et al., 2025; Ge et al., 2021), so detection-side metrics are not directly comparable across papers; we therefore report detection and association components separately and take the threshold-free AUROC of tracklet–detection similarity as our primary association-quality measure, which isolates the module under study from detector quality.

Metrics. End-to-end tracking is scored with TrackEval (Luiten & Hoffhues, 2020) (HOTA/DetA/AssA (Luiten et al., 2021), MOTA, IDF1, IDs, ID switches, fragments). Association quality is scored per epoch on the validation split by: the InfoNCE loss; the similarity classification accuracy at the learned threshold (sim ACC); and the AUROC of the tracklet–detection similarity scores, which is agnostic to any threshold choice. Energy and sparsity follow the protocol of Sec. 3.4.

Baselines and comparison axes. The evaluation is organized per dataset: Tables eftab:mot17–eftab:sportsmot report MOT17, DanceTrack, and SportsMOT independently, each comparing ASPC against the spiking counterpart (SNN) and published full-precision trackers on tracking accuracy (HOTA/AssA/DetA/IDF1/MOTA), activation sparsity, and per-frame energy. The *SNN baseline* is the same association architecture converted to leaky integrate-and-fire neurons with Hoyer-regularized thresholds (Datta et al., 2024; Neftci et al., 2019), trained identically from scratch on each dataset—the controlled, same-architecture spiking reference. Published full-precision trackers (MOTIP, MOTR, DiffMOT) are quoted from their original papers. Third-party spiking or sparse associators evaluated on the same splits are left for future comparison.

4.2 MAIN RESULTS: ASPC VS. SNN VS. ANN ON THREE DATASETS

Tables 1–3 organize the comparison per dataset, each benchmark held apart for independent third-party comparison. *Accuracy vs. SNN (same-recipe rows):* the identically configured spiking counterpart trails ASPC on every dataset—by 29.3 HOTA points on MOT17 (59.3 vs. 30.0), 22.5 on DanceTrack (61.2 vs. 38.7), and 19.8 on SportsMOT (75.5 vs. 55.7)—while costing up to $24\times$ more energy per frame; the spiking temporal accumulation fails to discriminate identities in a single pass, exactly the mismatch identified in Sec. 1. *Accuracy vs. full-precision trackers:* on MOT17 ASPC attains HOTA 59.3 / AssA 58.2, on par with the strongest full-precision baseline MOTIP (59.2 / 56.9) at 25.7 mJ against 1105 mJ; on DanceTrack ASPC (HOTA 61.2) trails only MOTIP (67.5) while exceeding DiffMOT (62.3) in AssA per joule at a fraction of the energy; on SportsMOT ASPC (HOTA 75.5, AssA 68.4) trails only DiffMOT's HOTA (76.2) while exceeding its AssA (65.1) and consuming 19 mJ against 557 mJ.

Cross-dataset robustness: the ternary representation does not sacrifice discriminative power despite 48–59% of its responses being gated to zero. The learning rule of Eq. 7 yields substantially different calibrated thresholds per dataset ($0.501/0.557/0.558$) that converge smoothly (Appendix C): the free-energy balance point is scene-specific and stable rather than a fixed constant. On MOT17

Table 1: MOT17 results. Sparsity: silent fraction of the gated units; Energy: per-frame inference energy (mJ); "n.r.": not reported.

Method	HOTA↑	AssA↑	DetA↑	IDF1↑	MOTA↑	Sparsity (%)	Energy (mJ/fr)
MOTIP (Gao et al., 2024)	59.2	56.9	60.2	**71.2**	**75.5**	n.r.	1105
MOTR (Zeng et al., 2022)	57.8	55.7	**60.3**	68.6	73.4	n.r.	n.r.
SNN (ours)	30.04	41.16	37.9	40.52	59.5	87.2	631
ASPC (ours)	**59.3**	**58.2**	60.2	68.5	72.6	59.2	**25.7**

Table 2: DanceTrack results. Sparsity: silent fraction of the gated units; Energy: per-frame inference energy (mJ); "n.r.": not reported.

Method	HOTA↑	AssA↑	DetA↑	IDF1↑	MOTA↑	Sparsity (%)	Energy (mJ/fr)
MOTIP (Gao et al., 2024)	**67.5**	**57.6**	79.4	**72.2**	90.3	n.r.	775
DiffMOT (Lv et al., 2024)	62.3	48.8	**82.5**	64.0	**92.7**	n.r.	651
SNN (ours)	38.73	19.66	76.89	36.86	63.64	87.6	103
ASPC (ours)	61.24	47.31	79.63	62.19	87.09	48.5	**11**

Table 3: SportsMOT results. Sparsity: silent fraction of the gated units; Energy: per-frame inference energy (mJ); "n.r.": not reported; the tracking scores are detection-limited (17 of 45 sequences yield no detections).

Method	HOTA↑	AssA↑	DetA↑	IDF1↑	MOTA↑	Sparsity (%)	Energy (mJ/fr)
MOTIP (Gao et al., 2024)	71.9	59.1	83.1	71.4	91.5	n.r.	745
DiffMOT (Lv et al., 2024)	**76.2**	65.1	**89.3**	76.1	**97.1**	dense	557
SNN (ours)	55.7	46.9	62.2	45.9	72.5	85.3	347
ASPC (ours)	75.5	**68.4**	79.9	**76.3**	91.1	48.7	**19**

identity management is tight—69 identities created against 75 ground-truth identities with only 65 switches—the prior-dominant regime of Sec. 3.3.

Energy vs. full-precision ANN and per-dataset inference energy. The ungated modules of the same association architecture, as run in CamelTrack and most tracker stacks (Wojke et al., 2017; Zhang et al., 2022b; Somers et al., 2025), execute 22.4G MACs, i.e., $\approx$ 103 mJ per frame—the dense reference; the energies of third-party trackers in Tables 1–3 are quoted from their original papers. ASPC's gated forward costs 79.7 mJ—a conservative $1.29\times$ reduction; with ternary operands exploited, each surviving multiply degenerates into a signed addition, pricing the 19.2G gated operations at ≈ 17 mJ. Per-frame energy is $25.7/11/19$ mJ on MOT17/DanceTrack/SportsMOT—$24\times$ below the SNN on MOT17 (631 mJ), $9.4\times$ on DanceTrack (103 mJ), and $18\times$ on SportsMOT (347 mJ). The zero-gating compresses the effective MAC count to 85.9–89.1% of the dense count, and the 2-bit ternary maps compress storage $16\times$—the operand-skipping pattern that event-driven accelerators are built to reap (Davies et al., 2018; Roy et al., 2019). Inference latency is 87.5–111.2 ms per frame in our unoptimized Python harness (including track-state bookkeeping); training consumes 112.8 kJ (0.66 h) on MOT17, 1212.6 kJ (13.4 h) on DanceTrack, and 1231.1 kJ (18.4 h) on SportsMOT—less than a single epoch of E2E trackers on 8 GPUs (Zeng et al., 2022; Gao & Wang, 2023).

Scene-adaptive sparsity. With one untuned recipe, the ternary networks allocate sparsity according to scene statistics (Tables 1–3): the silent fraction settles at 59.2% on MOT17's dense pedestrian scenes and $48.5\%/48.7\%$ on DanceTrack/SportsMOT, implemented automatically by the learned per-layer boundaries of Eq. 2; nothing is tuned per dataset.

Identity continuity under missing evidence. We probe the predictive module with two quantitative load-bearing behaviors. First, *recovery without re-spawn*: on the MOT17 held-out sequence ASPC creates 69 identities for 75 ground-truth identities and commits only 65 switches/fragmenting the track 54 times, i.e., once a person is known, the tracker tends to keep the identity alive through gaps instead of minting a new one—the prior-dominant regime of Sec. 3.3. Second, *association quality under detector failure*: on SportsMOT, where 17 of 45 validation sequences yield no detections at all,

Table 4: Activation-regime ablation on MOT17. Energy: per-frame inference energy (mJ).

Activation	λ_H	AUROC	sim ACC	zero (%)	Energy (mJ/fr)
SNN (LIF)	10^{-9}	0.9316	**0.9488**	87.2	631
SNN (LIF)	10^{-10}	0.9255	0.9196	86.3	647
Binary $\{0,1\}$	10^{-9}	0.9694	0.9307	**96.1**	**14.9**
Binary $\{0,1\}$	10^{-10}	0.9634	0.9361	95.8	15.8
ASPC (ours)	10^{-9}	**0.9798**	0.9411	59.2	25.7
ASPC (ours)	10^{-10}	0.9750	0.9357	58.6	26.1

ASPC still attains AssA 68.4 and IDF1 76.3 (Table 3) —association quality stays healthy even under a heavily degraded detector, because re-identification compares against the persistent internal state rather than against a leaked membrane. Under the identical recipe, the spiking baseline's HOTA on MOT17 is only 30.0 against ASPC's 59.3 (Table 1), reflecting exactly the failure mode of Sec. 1: the event-loss channel destroys the identity features the associator needs. DanceTrack and SportsMOT identity-switch counts remain dominated by detection errors (Tables 2 and 3), which we view as an instrumentation limit of the yolo11m-pose source rather than of the module.

4.3 Ablation studies

Activation regime. Table 4 isolates the activation mechanism on MOT17 under an otherwise identical recipe (random init, 30 epochs, Hoyer-regularized, from the same code base). The SNN is the regime ASPC displaces: 87.2% zeros but AUROC 0.932 and a 631 mJ/frame energy. The binary $\{0,1\}$ gate converges to extreme sparsity (96.1% zeros) with AUROC 0.969 at only 14.9 mJ per frame—showing that Hoyer-threshold gating itself transfers the representation power, but it codes only presence, discarding the sign that a metric-space embedder needs to express direction, and its AUROC trails the ternary gate by 1.0 point. The ternary ASPC restores sign at moderate sparsity (59.2% zeros, balanced $\pm 20\%$) with the highest AUROC 0.980 at 25.7 mJ measured—$24.5\times$ below the SNN's 631 mJ and within $1.7\times$ of the binary gate while adding the directional information the association needs. Loosening the Hoyer weight to 10^{-10} confirms the recipe's stability: the ternary gate holds AUROC 0.975 at 26.1 mJ.

Hoyer weight. On MOT17, sweeping λ_H by one decade (10^{-9} vs. 10^{-10}) moves the ternary gate's AUROC only from 0.980 to 0.975 and its energy from 25.7 to 26.1 mJ; the binary gate shows the same stability (0.969 vs. 0.963). The optimum sits at 10^{-9}: weaker regularization lets marginal units fire (slightly denser, no better AUROC), stronger regularization over-sparsifies and hurts discrimination. Since the regularizer's force is geometric (Sec. 3.4), even the winning weight's numerical loss contribution to the objective is negligible ($\approx 5\times 10^{-7}$)— it works by shaping the pre-activation distribution around the gate boundary.

5 Conclusion

We presented ASPC, an online tracking framework that combines two brain-inspired principles: sparse signed coding, realized as a ternary activation network with Hoyer-extremum gating, and prediction-driven association, realized as free-energy minimization over persistent internal track states. The combination relocates sparsity from the time domain to the feature space and turns the matching threshold from a heuristic into a learned balance point. On MOT17, DanceTrack, and SportsMOT, ASPC matches or surpasses the strongest full-precision trackers in HOTA and AssA while far exceeding its identically configured SNN counterpart and consuming per-frame energies of 25.7 mJ, 11 mJ, and 19 mJ—well below both the dense FP32 regime and the spiking regime. We believe this operating point opens a practical path to continuous, low-power multi-object perception; promising directions include co-design with ternary-capable accelerators and extending gating to the detector itself.

REPRODUCIBILITY STATEMENT

We report the full training recipe (activation gating, Hoyer weights, epochs, initialization, and matching-threshold calibration) in Secs. 3.4 and 4.1, the energy measurement protocols and constants in Sec. 3.4 and Appendix B, and the per-epoch learned thresholds and reproduction commands in Appendix C. All experiments use publicly available datasets (MOT17, DanceTrack, SportsMOT) with splits stated in Sec. 4.1; code and configuration files accompany the submission.

AI USE STATEMENT

We used large language models as writing aids to polish grammar and phrasing of the manuscript and to assist with drafting figure captions; all technical content, experimental design, analysis, and conclusions are the authors' own and were verified by the authors. No AI tools were used to generate experimental results, data, or code beyond standard library usage. The authors reviewed and edited all AI-assisted output and take full responsibility for the content of this paper.

REFERENCES

Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. BoT-SORT: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022.

Patrick Bergmann, Timm Meinhardt, and Laura Leal-Taixé. Tracking without bells and whistles. In *ICCV*, 2019.

Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *ICIP*, 2016.

Tongbu Bu, Wei Fang, Jianhao Ding, Pengjun Dai, Zhaofei Yu, and Tiejun Huang. Optimal ANN-SNN conversion for high-accuracy and ultra-low-latency spiking neural networks. In *ICLR*, 2022.

Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric SORT: Rethinking SORT for robust multi-object tracking. In *CVPR*, 2023.

Hanting Chen, Yunhe Wang, Chao Xu, Boxin Shi, Chijin Xu, Qi Tian, and Chang Xu. AdderNet: Do we really need multiplications in deep learning? In *CVPR*, 2020.

Sayeed Shafayet Chowdhury, Nitin Rathi, and Kaushik Roy. One timestep is all you need: Training spiking neural networks with ultra low latency. *arXiv preprint arXiv:2110.05929*, 2021.

Pengfei Chu, Jinwang Wang, Quan You, Haibin Ling, and Zhicheng Liu. TransMOT: Spatial-temporal graph transformer for multiple object tracking. In *WACV*, 2023.

Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or -1. *arXiv preprint arXiv:1602.02830*, 2016.

Yutao Cui, Chengjian Zeng, Xiaoyang Zhao, Yang Yang, Guosheng Wu, and Limin Wang. SportsMOT: A large multi-object tracking dataset in multiple sports scenes. In *ICCV*, 2023.

Gourav Datta, Zhuang Liu, and Peter A. Beerel. Can we get the best of both binary neural networks and spiking neural networks for efficient computer vision? In *ICLR*, 2024.

Howard David, Eugene Gorbatov, Ulf R. Hanebutte, Rahul Khanna, and Christian Le. RAPL: Memory power estimation and capping. In *ISLPED*, 2010.

Mike Davies et al. Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro*, 2018.

Shikuang Deng, Yuhang Li, Shanghang Zhang, and Shi Gu. Temporal efficient training of spiking neural network via gradient re-weighting. In *ICLR*, 2022.

Simei Deng and Yu Gu. Optimal conversion of conventional artificial neural networks to spiking neural networks. In *ICLR*, 2021.

Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11: 127–138, 2010.

Karl Friston, Francesco Rigoli, Dimitri Ognibene, Christoph Mathys, Thomas Fitzgerald, and Giovanni Pezzulo. Active inference and epistemic value. *Cognitive Neuroscience*, 6(4):187–214, 2015.

Rui Gao and Lijun Wang. MeMOTR: Long-term memory-augmented transformer for multi-object tracking. In *ICCV*, 2023.

Rui Gao, Yifu Zhang, and Lijun Wang. Multiple object tracking as ID prediction. *arXiv preprint arXiv:2403.16848*, 2024.

Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jie Sun. YOLOX: Exceeding YOLO series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

Mark Horowitz. 1.1 Computing's energy problem (and what we can do about it). In *ISSCC*, 2014.

Patrik O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.

Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. In *NeurIPS*, 2016.

Tao Jiang, Ping Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. RTMPose: Real-time multi-person pose estimation based on MMPose. *arXiv preprint arXiv:2303.07399*, 2023.

Glenn Jocher and Jing Qiu. Ultralytics YOLO11. `https://github.com/ultralytics/ultralytics`, 2025.

Rudolf E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82:35–45, 1960.

Seijoon Kim, Seongsik Park, Byunggook Na, and Sungroh Yoon. Spiking-YOLO: Spiking neural network for energy-efficient object detection. In *AAAI*, 2020.

Harold W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.

Mark Kurtz, Justin Kopinsky, Rati Gelashvili, Alexander Matveev, John Carr, Michael Goin, William Leiserson, Sage Moore, Nir Shavit, and Dan Alistarh. Inducing and exploiting activation sparsity for fast inference on deep neural networks. In *ICML*, 2020.

Zechun Liu, Baoyuan Wu, Wenhan Luo, Xin Yang, Wei Liu, and Kwang-Ting Cheng. Bi-Real Net: Enhancing the performance of 1-bit CNNs with improved representational capability and advanced training algorithm. In *ECCV*, 2018.

Zechun Liu, Zhiqiang Shen, Marios Savvides, and Kwang-Ting Cheng. ReActNet: Towards precise binary neural network with generalized activation functions. In *ECCV*, 2020.

Jonathon Luiten and Arne Hoffhues. TrackEval. `https://github.com/JonathonLuiten/TrackEval`, 2020.

Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. HOTA: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, 2021.

Weitao Lv, Ying Huang, Naiyan Zhang, Richard-She-uei Lin, Minghan Han, and Dylan Zeng. Diff-MOT: A real-time diffusion-based multiple object tracker with non-linear prediction. In *CVPR*, 2024.

Gerald Maggiolino, Ahmad Ahmad, Jinkun Cao, and Kris Kitani. Deep OC-SORT: Multi-pedestrian tracking by adaptive re-identification. In *ICIP*, 2023.

Timm Meinhardt, Alexander Kirillov, Laura Leal-Taixé, and Christoph Feichtenhofer. TrackFormer: Multi-object tracking with transformers. In *CVPR*, 2022.

Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. MOT16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016.

Emre O. Neftci, Hesham Mostafa, and Friedemann Zenke. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36(6):51–63, 2019.

Wouter W. L. Nuijten, Mariia Lukashchuk, Thijs van de Laar, and Bert de Vries. A message passing realization of expected free energy minimization. *arXiv preprint arXiv:2508.02197*, 2026.

NVIDIA Corporation. NVIDIA management library (NVML). `https://developer.nvidia.com/nvidia-management-library-nvml`, 2025.

Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

Ben Poole, Sherjil Ozair, Aäron van den Oord, Alexander A. Alemi, and George Tucker. On variational bounds of mutual information. In *ICML*, 2019.

Antonis Psalta, Vasileios Tsironis, and Konstantinos Karantzalos. Transformer-based assignment decision network for multiple object tracking. *Computer Vision and Image Understanding*, 241: 103957, 2024.

Zhiyin Qin, Shijian Zhou, Linkai Wang, Jiangning Duan, Gang Hua, and Wei Tang. MotionTrack: Learning robust short-term and long-term motions for multi-object tracking. In *CVPR*, 2023.

Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999.

Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. XNOR-Net: ImageNet classification using binary convolutional neural networks. In *ECCV*, 2016.

Nitin Rathi, Gopalakrishnan Srinivasan, Priyadarshini Panda, and Kaushik Roy. Enabling deep spiking neural networks with hybrid conversion and spike timing dependent backpropagation. *arXiv preprint arXiv:2005.01807*, 2020.

Kaushik Roy, Akhilesh Jaiswal, and Priyadarshini Panda. Towards spike-based machine intelligence with neuromorphic computing. *Nature*, 575:607–617, 2019.

Charbel Sakr, Junsin Choi, Zhe Wang, Kailash Gopalakrishnan, and Naresh Shanbhag. True gradient-based training of deep binary activated neural networks via continuous binarization. *IEEE Transactions on Neural Networks and Learning Systems*, 2019.

Jens Seidenschwarz, Guillem Brasó, Victor Castro Serrano, Ismail Elezi, and Laura Leal-Taixé. Simple cues lead to a strong multi-object tracker. In *CVPR*, 2023.

Vladimir Somers, Bram Standaert, Victor Joos, Alexandre Alahi, and Christophe De Vleeschouwer. CAMELTrack: Context-aware multi-cue exploitation for online multi-object tracking. *arXiv preprint arXiv:2505.01257*, 2025.

Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. DanceTrack: Multi-object tracking in uniform appearance and diverse motion. In *CVPR*, 2022.

Zheng Tu, Xinhai Chen, Pengfei Ren, and Yadong Wang. AdaBin: Improving binary neural networks with adaptive binary sets. In *ECCV*, 2022.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ashish Vaswani et al. Attention is all you need. In *NeurIPS*, 2017.

Peisong Wang, Xueyu He, Gang Li, Tianli Zhao, and Jian Cheng. Sparsity-inducing binarized neural networks. In *AAAI*, 2020.

Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, 2017.

Huanrui Yang, Wei Wen, and Hai Li. DeepHoyer: Learning sparser neural network with differentiable scale-invariant sparsity measures. In *ICLR*, 2020.

Minghao Yang, Guolei Han, Bingxin Yan, Wenzhe Zhang, Junjie Qi, Huihui Lu, and Dong Wang. Hybrid-SORT: Weak cues matter for online multi-object tracking. In *AAAI*, 2024.

Shuangming Yang and Badong Chen. Effective surrogate gradient learning with high-order information bottleneck for spike-based machine intelligence. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

Shuangming Yang, Bernab'e Linares-Barranco, Yuzhu Wu, and Badong Chen. Self-supervised high-order information bottleneck learning of spiking neural network for robust event-based optical flow estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Shuangming Yang, Jianshi Tang, Tianhang Zhou, Bernab'e Linares-Barranco, Giacomo Indiveri, Xinhu Zheng, Mostafa Rahimi Azghadi, and Badong Chen. Efficient training of neuromorphic electronics. *Nature Electronics*, 2026.

Fangao Zeng, Bin Dong, Tiancai Zhang, Tianheng Wang, Xiangyu Zhang, and Yichen Wei. MOTR: End-to-end multiple-object tracking with transformer. In *ECCV*, 2022.

Yichi Zhang, Zhi Zhang, and Lukasz Lew. PokeBNN: A binary pursuit of lightweight accuracy. In *CVPR*, 2022a.

Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. FairMOT: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 2021.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-object tracking by associating every detection box. In *ECCV*, 2022b.

Yifu Zhang, Tiancai Wang, and Xiangyu Zhang. MOTRv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors. In *CVPR*, 2023.

Hanle Zheng, Yujie Wu, Lei Deng, Yifan Hu, and Guoqi Li. Going deeper with directly-trained larger spiking neural networks. In *AAAI*, 2021.

A Appendix A: Ternary gate derivations

Lemma 1Hoyer extremum. *Let $m \in \mathbb{R}_+^n$ be a non-negative tensor with $\|m\|_1 > 0$. The Hoyer sparseness measure $\mathcal{H}(m) = \left(\|m\|_1/\|m\|_2\right)^2$ is scale-invariant, and among all vectors with fixed $\|m\|_1$ and $\|m\|_2$, $\mathcal{H}$ attains its minimum on the set of vectors whose positive entries are all equal to $m^\star = \|m\|_2^2/\|m\|_1$.*

Proof. Scale invariance is immediate since both $\|m\|_1$ and $\|m\|_2$ are homogeneous of degree one. For the extremum, fix $\|m\|_2$; minimizing $\mathcal{H}$ is equivalent to minimizing $\|m\|_1$ subject to $\|m\|_2$ fixed and $m \geq 0$. By Cauchy–Schwarz, $\|m\|_1 \geq \|m\|_2^2/\|m\|_\infty$, with equality exactly when all positive entries equal $\|m\|_\infty$; substituting the extremal value gives $m^\star = \|m\|_2^2/\|m\|_1$ as the marginal-mass boundary, following (Datta et al., 2024; Hoyer, 2004). □

Proposition 1Gradient of the Hoyer measure. *$\mathcal{H}$ is differentiable almost everywhere, with $\partial\mathcal{H}/\partial m \propto \|m\|_1 \operatorname{sign}(m) - \|m\|_1^2\, m/\|m\|_2^2$; the gradient pushes magnitudes away from $m^\star$, displacing the activation population from the gate boundary.*

Hoyer extremum. The Hoyer sparseness measure $\mathcal{H}(m) = \left(\|m\|_1/\|m\|_2\right)^2$ on a non-negative tensor m is scale-invariant and differentiable almost everywhere, with gradient $\partial\mathcal{H}/\partial m \propto \|m\|_1 \operatorname{sign}(m) - \|m\|_1^2\, m/\|m\|_2^2$; setting it to zero for positive entries gives the extremum $m^\star = \|m\|_2^2/\|m\|_1$, which for non-negative m satisfies $m^\star \le \max_j m_j$ and is the minimum of $\mathcal{H}$ (second derivative positive), following (Datta et al., 2024). Using the magnitude map $m_l = |z_l|$ makes the extremum a symmetric magnitude boundary, which the gate of Eq. 2 applies to both signs—hence the $\{-1, 0, +1\}$ output.

Surrogate gradient. The gate's derivative is zero almost everywhere and undefined at the boundaries; Eq. 3 in the main text adopts the box surrogate of BANN (Datta et al., 2024) symmetric around both boundaries $\pm E_l$, i.e., gradient flows only to units whose magnitude lies within E_l of the boundary, preventing the silent (zero) region from receiving updates while keeping the signed regions trainable via the straight-through estimator.

B Appendix B: Energy measurement protocol

Layer-wise model. Per-frame forward energy $E = \sum_l \left[(1-s_l)\,\mathrm{MACs}_l \cdot 4.6\,\mathrm{pJ} + \mathrm{ACs}_l \cdot 0.9\,\mathrm{pJ}\right]$ using 45 nm constants (FP32 MAC 4.6 pJ, accumulate 0.9 pJ) (Horowitz, 2014), where s_l is the layer's zero-activation ratio, i.e., MACs are charged only on non-zero (ternary) activations; the backward pass is estimated with a factor of 2.0 over the forward counters, and attention softmax/normalization layers are counted dense (they admit no operand gating in our realization), which is why effective MACs remain 85.9–89.1% of the dense total rather than tracking the zero ratio one-to-one. The SNN estimate uses the same constants with spike-driven MAC/AC counting per the standard model (Datta et al., 2024).

Hardware attribution. A background thread samples GPU power at card granularity (NVML (NVIDIA Corporation, 2025)) and CPU package/DRAM power (RAPL (David et al., 2010)) at $\sim$1 Hz for the entire run. An idle baseline is captured on the same GPU before the run; attributable energy is (stage total) $-$ (idle baseline $\times$ stage duration). Per-frame inference energy is the attributable energy of the final validation stage divided by the frames processed in that stage (100 frames on MOT17, 2496 on DanceTrack, 4491 on SportsMOT). Training energy is the attributable energy of the full training stage. Latency is wall-clock stage duration per frame under the same conditions.

C Appendix C: Learned thresholds and reproduction

Per-epoch learned matching thresholds. The threshold of Eq. 7 is calibrated per epoch on the validation split; its evolution shows smooth convergence to a dataset-specific balance point:

MOT17: `0.553 0.554 0.551 0.517 0.484 0.526 0.567 0.575 0.532 0.545 0.553 0.539 0.557 0.507 0.495 0.520 0.551 0.516 0.474 0.482 0.502 0.497 0.474 0.459 0.498 0.461 0.499 0.499 0.494 0.488 0.501`

DanceTrack: `0.550 0.516 0.503 0.494 0.498 0.495 0.488 0.486 0.484 0.492 0.489 0.513 0.513 0.486 0.539 0.521 0.518 0.543 0.551 0.558 0.564 0.571 0.554 0.558 0.550 0.567 0.544 0.550 0.572 0.553 0.557`

SportsMOT: `0.571 0.535 0.535 0.552 0.562 0.572 0.566 0.536 0.556 0.597 0.557 0.591 0.552 0.557 0.570 0.579 0.608 0.567 0.571 0.585 0.550 0.592 0.568 0.563 0.551 0.560 0.563 0.561 0.565 0.556 0.558`

Reproduction recipe. All models are trained from scratch (random initialization), 30 epochs, λ_H=10^{-9}, per dataset: (1) pre-generate tracker states (detections + ReID + pose) with `yolo11m-pose` (SportsMOT: 1920-px inference, confidence 0.05, 0.1 floor) using the tracklab pipeline; (2) train from scratch:

```
tracklab -cn cameltrack_train dataset=<ds> \
  modules.track.checkpoint_path=null \
  modules.track.CAMEL.hoyer_lambda=1e-9
```

with InfoNCE on cross-video association scenarios and identity-swap, detection-dropout, and cue-dropout augmentations; (3) evaluate on the official validation split with TrackEval using the calibrated threshold.