CDMMT: CONDITIONAL DIFFUSION WITH MULTI-MODAL GUIDANCE FOR MULTI-OBJECT TRACKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Because data modalities differ substantially, existing multimodal tracking methods often design specialized mechanisms for specific modalities, making it difficult to use multimodal information effectively within a unified framework. To address this issue, we propose Conditional Diffusion for Multimodal Multi-Object Tracking (CDMMT). Through modality-adapted condition encoding and injection, CDMMT incorporates different forms of auxiliary information into a shared paired-box denoising process for object localization and adjacent-frame association. Modality information is first encoded as a condition and injected into paired-box prediction. We further introduce energy-guided sampling: differentiable conditional energies measure the agreement between predicted boxes and modality information, and their gradients correct intermediate box predictions so that the constraints propagate to later predictions through sampling-state updates. Using Global Navigation Satellite System (GNSS) positions and infrared images, we demonstrate the encoding, injection, and guidance of sparse positional and dense image information. Experiments show that CDMMT effectively uses multiple modalities and improves tracking performance. Code is available anonymously at `https://anonymous.4open.science/r/CDT-B163`.

1 INTRODUCTION

Multimodal tracking aims to improve robustness by using information beyond RGB vision, including infrared images, depth, and satellite positioning. These modalities can provide complementary constraints. Existing methods, however, usually design information-use mechanisms for particular modalities, such as feature fusion for infrared images or association strategies for depth. Modalities differ markedly in data form, spatial granularity, and semantic meaning. This work asks how to preserve those differences while allowing each modality to participate in a shared object-state estimation process.

We propose Conditional Diffusion for Multimodal Multi-Object Tracking (CDMMT) to address this problem. Multi-object tracking requires locating objects in each frame and establishing correspondence across frames. We use a candidate box pair from two adjacent frames as the prediction unit. Each pair contains one box in the reference frame and one in the current frame, representing two positions and their potential correspondence. Starting from noisy box coordinates, the model combines RGB features and auxiliary modality conditions to jointly predict the two object boxes and their association confidence. Trajectory matching then maintains object identities. Thus, box generation predicts object positions, while paired-box prediction provides cross-frame correspondences.

Auxiliary modalities participate in prediction through interfaces adapted to their data forms. Positioning information is encoded into condition vectors and updates candidate features through cross-attention. Infrared images are encoded into multiscale features that augment RGB features before regional feature extraction. Both routes affect the predictions of the paired-box estimator and therefore enter the reverse denoising process. They share the paired-box representation, prediction procedure, and sampling rule while retaining their own condition encoders and fusion methods.

We additionally introduce energy-guided sampling. A differentiable conditional energy describes the mismatch between paired-box predictions and modality information. At intermediate denoising steps, one gradient correction moves predicted boxes toward lower energy; the corrected predictions then determine the DDIM state update. The updated candidate state enters the next network call,

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.

which re-estimates boxes, class scores, and association information. We evaluate tracking under GNSS positioning and infrared-image conditions. Positioning provides sparse object-level information, whereas infrared images provide thermal radiation information complementary to visible appearance. A depth-modality extension appears in the supplementary material.

Our contributions are as follows:

- We propose CDMMT, which formulates box generation and cross-frame association as paired-box denoising guided by different modality conditions, incorporating heterogeneous information into a shared box-state representation, estimator, and sampling process.
- We develop condition-adaptation methods for GNSS positioning and infrared images, introducing sparse object-level and dense image-level conditions, respectively, and validate them in two multimodal multi-object tracking settings.

2 Related Work

2.1 Multi-Object Tracking

Tracking by detection. Tracking-by-detection methods usually detect objects frame by frame and then associate them using motion, appearance, or geometric cues. SORT (Bewley et al., 2016) uses a Kalman filter and Hungarian matching for efficient online tracking, but prediction errors can accumulate under occlusion and nonlinear motion. DeepSORT (Wojke et al., 2017) adds ReID appearance features to help restore identities, although low resolution and similar appearances remain difficult. ByteTrack (Zhang et al., 2022) associates low-confidence detections to reduce track fragmentation from missed detections. OC-SORT (Cao et al., 2023), GHOST (Seidenschwarz et al., 2023), and Hybrid-SORT (Yang et al., 2024) improve association robustness through observation-based state estimation, simple appearance cues, and fusion of weak cues, respectively. MotionTrack (Qin et al., 2023) models short- and long-term motion; FineTrack (Ren et al., 2023) strengthens appearance discrimination with fine-grained representations; and DeconfuseTrack (Huang et al., 2024) designs decomposed data association for confusion in detection, trajectories, and association. TrackTrack (Shim et al., 2025) shifts attention from single-frame detection to trajectory representation and trajectory-level association, while SAM2MOT (Jiang et al., 2026a) explores tracking by segmentation using SAM2.

End-to-end multi-object tracking. End-to-end approaches jointly learn detection, state propagation, and identity association in one network. TrackFormer (Meinhardt et al., 2022) propagates objects through object and track queries in a Transformer. MOTR (Zeng et al., 2022) introduces track queries and tracklet-aware label assignment to update instances frame by frame. MOTRv2 (Zhang et al., 2023) uses anchor-form queries, taking proposals from an additionally pretrained detector as query anchors to supply detection priors and ease the conflict between detection and association learning. MeMOT (Cai et al., 2022) maintains spatiotemporal memory of identity representations and adaptively aggregates historical information. MeMOTR (Gao & Wang, 2023) injects long-term memory into track embeddings through memory attention. GTR (Zhou et al., 2022) encodes object features from all frames of a video clip and aggregates them into global trajectories using trajectory queries. MOTIP (Gao et al., 2025) casts cross-frame association as ID prediction and updates trajectories by directly predicting object identities.

Diffusion-based multi-object tracking. The progressive recovery of object states from noise offers an iterative way to model box localization, candidate propagation, and motion prediction. DiffusionTrack (Luo et al., 2024) jointly casts detection and association as denoising paired noisy boxes, establishing a generative tracking paradigm. Similarly, Pro2Diff (Liu et al., 2024) formulates tracking as diffusion-based proposal propagation that gradually refines object states. ConsistencyTrack (Jiang et al., 2026b) replaces multistep reversal on paired noisy boxes with a self-consistent mapping for one- or few-step generation and adds multistage association for occluded objects. DiffMOT (Lv et al., 2024) models nonlinear dynamics by predicting box displacement with diffusion in place of a conventional Kalman filter, while largely retaining standard detection and association. DiffusionMOT (Hu et al., 2025) further combines paired-box diffusion with diffusion-based detection and motion prediction. Overall, existing diffusion trackers mainly condition denoising on visual information or internal tracking states; external modalities as conditions for paired-box denoising remain underexplored.

2.2 MULTIMODAL TRACKING

Multimodal tracking introduces infrared, depth, or radar information to compensate for RGB limitations under low light, occlusion, and similar appearances. RGB-T tracking uses thermal responses to strengthen visible-light appearance representations. MANet (Li et al., 2019) analyzes pixel-, feature-, and response-level fusion and shows that fusion location affects modality complementarity. MFGNet (Wang et al., 2023) uses dynamic modality-aware filters to adaptively balance visible and infrared branches and reduce interference from a low-quality modality. MACFT (Luo et al., 2023) strengthens cross-modal interaction with mixed attention. For RGB-T tracking of multiple small objects, HGT-Track (Xu et al., 2024) uses a heterogeneous graph Transformer to aggregate spatial and temporal relations into cross-modal features. RGB-D methods use depth to represent relative object ordering. DP-MOT (Quach et al., 2024) establishes depth order in occlusions with subject-ordered depth estimation and introduces depth-aware Kalman updates. SparseTrack (Liu et al., 2025) divides crowded scenes into sparse subsets using pseudo-depth to reduce matching ambiguity. Radar methods additionally supply geometric distance and velocity. CenterFusion (Nabati & Qi, 2021) associates radar points with image object centers through frustum-based association to improve camera-based 3D detection. CFTrack (Nabati et al., 2021) combines camera-radar fusion detections with association based on depth, velocity, and 2D displacement for robust 3D tracking. We incorporate sparse object-level and dense image-level conditions into a shared paired-box denoising process while preserving their distinct information structures.

3 PROPOSED METHOD

3.1 CONDITIONAL DIFFUSION FOR MULTIMODAL TRACKING

CDMMT uses candidate box pairs from adjacent frames as prediction units. It denoises them using RGB features and auxiliary modality conditions, then maintains identities by trajectory matching. Given images I_{t-1} and I_t, a visual encoder extracts F_{t-1} and F_t. Let M denote auxiliary modality information available to this prediction and $C = \mathrm{Enc}_m(M)$ its conditional representation. We write M_{loc} and M_{ir} for GNSS positioning and infrared-image information. Each modality has its own encoding and injection method, while paired-box estimation and sampling are shared. During inference, conditional energies additionally correct intermediate box predictions and affect later denoising through state updates. We use t for the video-frame index and s for the diffusion noise level.

Paired-box representation and forward diffusion. As shown in Fig. 1, training first aligns ground-truth boxes in adjacent frames by trajectory identity. Boxes of the same object form a ground-truth pair. If their number exceeds the prescribed candidate count N, pairs are jointly subsampled using the same pairing indices; otherwise random box pairs are added.

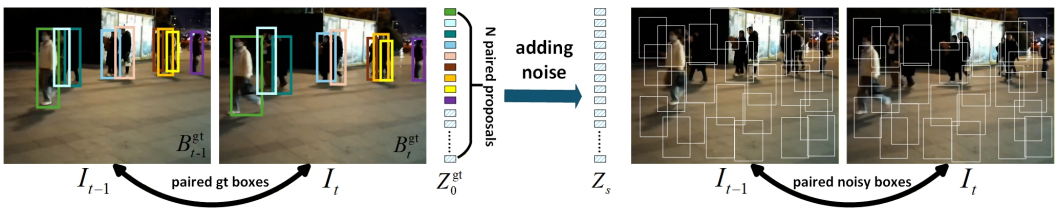


Figure 1: Paired-box construction and forward diffusion. Ground-truth boxes are aligned by trajectory identity; paired boxes are jointly subsampled or randomly padded to form Z_0^{gt}, and Gaussian noise is added to obtain Z_s. Matching colors indicate corresponding boxes. Solid and hatched fills denote ground-truth and randomly padded pairs, respectively; the right side shows noisy boxes.

For candidate i in frame $\tau \in \{t-1, t\}$, let $b_{\tau,i} = (c_x, c_y, w, h)$ denote normalized center coordinates, width, and height. Boxes with the same index in the two frames form a candidate pair. Arranging

N pairs as rows gives the uncorrupted representation

$$z_{0,i}^{\mathrm{gt}} = [b_{t-1,i}, b_{t,i}] \in \mathbb{R}^8, \qquad Z_0^{\mathrm{gt}} = \begin{bmatrix} z_{0,1}^{\mathrm{gt}} \\ \vdots \\ z_{0,N}^{\mathrm{gt}} \end{bmatrix} \in \mathbb{R}^{N\times 8}. \tag{1}$$

During training, a noise level s is sampled and Gaussian noise is added to paired-box coordinates according to the diffusion process (Ho et al., 2020):

$$Z_s = \sqrt{\bar{\alpha}_s} Z_0^{\mathrm{gt}} + \sqrt{1-\bar{\alpha}_s}\epsilon, \qquad \epsilon \sim \mathcal{N}(0, \mathbf{I}). \tag{2}$$

Here $\bar{\alpha}_s = \prod_{u=1}^{s}(1-\beta_u)$; β_u is the noise variance from a cosine schedule, and $\bar{\alpha}_s$ controls the retained proportion of the original coordinate signal. Noise levels are $s \in \{1, \ldots, T\}$, with 0 denoting the uncorrupted state. Coordinate perturbation preserves candidate-pair indices.

Conditional paired-box estimation. The reverse process predicts object positions from the current noisy boxes, RGB features, and auxiliary condition, then updates the noisy state. Given C, the conditional transition from noise level s to $s' < s$ is modeled as follows (fixed RGB features are omitted from the probability expression for brevity):

$$p_\theta(Z_{s'} \mid Z_s, C) = \mathcal{N}\big(\mu_\theta(Z_s, s, C), \sigma_{s,s'}^2 \mathbf{I}\big)\,. \tag{3}$$

Here μ_θ is the transition mean and $\sigma_{s,s'}$ its sampling standard deviation. We use a clean-box prediction parameterization. At each DDIM (Song et al., 2021) step, the network predicts from the current state, RGB features, and condition:

$$\hat{Z}_s^{\mathrm{c}} = f_\theta(Z_s, s, F_{t-1}, F_t; C)\,. \tag{4}$$

$\hat{Z}_s^{\mathrm{c}}$ denotes the clean-box estimate from state s, with superscript c indicating a conditional prediction. Subsequent noise reconstruction uses this prediction to determine the sampling-state update.

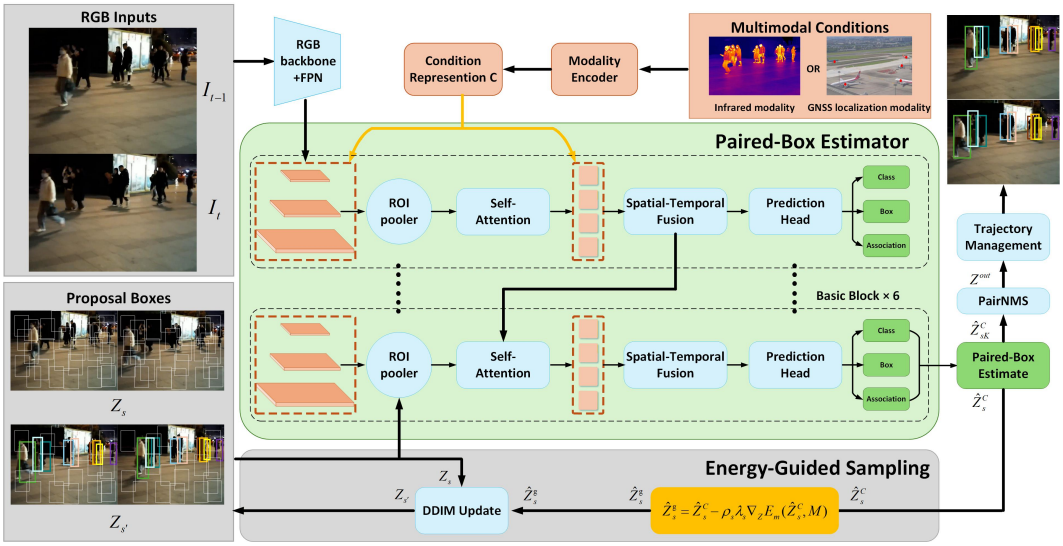


Figure 2: Conditional denoising, energy-guided sampling, and trajectory management. Yellow paths indicate the injection locations for different modality conditions. Energy-corrected intermediate box predictions participate in DDIM updates; the final step follows the upper path to PairNMS. Trajectory management maintains video-level identities from the filtered paired boxes.

As shown in Fig. 2, one call to f_θ contains six cascaded Basic Blocks. The first block initializes proposal representations by spatial pooling of ROI features from the two feature pyramids. Each subsequent block inherits those representations and extracts new ROI features at its updated boxes. Self-attention exchanges information among proposals within a frame, whereas cross-frame fusion incorporates the paired region in the other frame. Three heads then predict framewise class scores,

box increments, and a shared association confidence. Passing updated boxes and features to the next block progressively refines both localization and correspondence.

GNSS conditions update candidate features after self-attention through cross-attention; infrared conditions enter the feature pyramid as multiscale residuals before ROI extraction. The modality sections below define these two routes.

Energy-guided sampling. Feature injection affects the network prediction internally; energy guidance imposes an explicit condition in box-coordinate space. Let differentiable $E_m(Z, M)$ measure disagreement between a set of paired boxes and auxiliary modality information, so lower energy means greater agreement; m denotes the modality. At level s, define a local Gaussian proxy centered on the conditional estimate, $\widetilde{p}_s(Z) = \mathcal{N}(\mathrm{vec}(Z); \mathrm{vec}(\hat{Z}_s^{\mathrm{c}}), \tau_s^2 \mathbf{I})$, where $\tau_s > 0$ is a local scale and vec vectorizes a matrix. Reweight this proxy by the energy:

$$\widetilde{q}_s(Z) = \frac{1}{A_s} \widetilde{p}_s(Z) \exp[-\lambda_s E_m(Z, M)]. \tag{5}$$

A_s is a normalizing constant, and $\lambda_s \geq 0$ controls the energy weight. The local proxy is distinct from the learned conditional transition p_θ. Reweighting increases the relative density of low-energy box sets, combining network predictions and modality constraints locally. Maximizing the posterior of this proxy is equivalent to minimizing its negative log density:

$$Z_s^\star = \arg\min_Z \mathcal{J}_s(Z), \qquad \mathcal{J}_s(Z) = \frac{\|Z - \hat{Z}_s^{\mathrm{c}}\|_F^2}{2\tau_s^2} + \lambda_s E_m(Z, M). \tag{6}$$

The first term limits deviation from the conditional prediction, while the second encourages modality agreement. Starting from $Z^{(0)} = \hat{Z}_s^{\mathrm{c}}$, one gradient-descent step approximates this local MAP solution. Because the fidelity gradient is zero at that point,

$$\hat{Z}_s^{\mathrm{g}} = \hat{Z}_s^{\mathrm{c}} - \rho_s \lambda_s \, \nabla_Z E_m(Z, M)|_{Z=\hat{Z}_s^{\mathrm{c}}} . \tag{7}$$

Here $\rho_s > 0$ is the step size, $\lambda_s \geq 0$ the energy weight, and g denotes the guided estimate. The correction changes only box coordinates, not network parameters; τ_s does not explicitly enter this one-step update because the fidelity gradient vanishes at initialization.

DDIM update. Take K descending network-call levels $\mathcal{S} = \{s_1 > \cdots > s_K \geq 1\}$. Only when $k < K$ is the conditional prediction corrected by the energy; the corrected clean boxes then reconstruct the noise. For $s = s_k$ and $s' = s_{k+1}$, the update is

$$\begin{aligned} \hat{\epsilon}_s^{\mathrm{g}} &= \frac{Z_s - \sqrt{\bar{\alpha}_s}\hat{Z}_s^{\mathrm{g}}}{\sqrt{1 - \bar{\alpha}_s}}, \\ Z_{s'} &= \sqrt{\bar{\alpha}_{s'}}\hat{Z}_s^{\mathrm{g}} + \sqrt{1 - \bar{\alpha}_{s'} - \sigma_{s,s'}^2}\, \hat{\epsilon}_s^{\mathrm{g}} + \sigma_{s,s'}\xi, \qquad \xi \sim \mathcal{N}(0, \mathbf{I}). \end{aligned} \tag{8}$$

Here $\sigma_{s,s'}$ is the DDIM sampling-noise standard deviation; its full form is in the supplementary material. Setting $\eta_{\mathrm{DDIM}} = 0$ gives a deterministic update, while $\lambda_s = 0$ recovers unguided conditional sampling. The next level s' may skip intermediate noise levels.

Trajectory generation. Inference begins from Gaussian random candidates; when tracks exist, historical and additional random boxes can also be noised for initialization. Each updated $Z_{s'}$ enters another network call that re-estimates boxes, class scores, and association confidence. The final call applies neither energy correction nor a DDIM update, keeping its final boxes and scores aligned. Predictions are converted to image coordinates, filtered by class and association scores, and processed by paired-box NMS, abbreviated as

$$Z^{\mathrm{out}} \triangleq \mathrm{PairNMS}\left(\hat{Z}_{s_K}^{\mathrm{c}}\right). \tag{9}$$

Class and association score arguments are omitted. For each retained pair, the tracker matches the reference-frame box to an existing track and propagates the matched track ID to the current-frame box along the pair correspondence. High-confidence current-frame detections that do not continue an existing track start new tracks. Tracks without a current match retain their historical identities for later reassociation. The network predicts pairs and scores; the inference-time trajectory manager maintains IDs. Further details are in the supplementary material.

Loss function. During training, set matching assigns predicted box pairs to ground-truth pairs; randomly padded boxes construct inputs but are not ground-truth annotations. For predicted pair T_d, joint classification confidence is $\widetilde{\pi}_d^\tau = \sqrt{\hat{\pi}_d^\tau \hat{a}_d}$, where $\hat{\pi}_d^\tau$ is the class probability in frame τ and $\hat{a}_d$ is association confidence. Matching combines this confidence with two-frame box L_1 distances and paired-box GIoU to obtain the matched set $\mathcal{A}$. Matched candidates receive classification and box-regression supervision; unmatched candidates are classification background. The overall loss is

$$\mathcal{L}_{pair} = \lambda_1 \mathcal{L}_{cls} + \frac{1}{N_{pos}} \sum_{(T_d, T_{gt}) \in \mathcal{A}} \left[\lambda_2 \mathcal{L}_{reg}(T_d, T_{gt}) + \lambda_3 \big(1 - \mathrm{GIoU}_{3D}(T_d, T_{gt})\big) \right]. \tag{10}$$

Here $\mathcal{L}_{cls}$ is Focal Loss over all candidates in both frames, normalized by the number of positives; $\mathcal{L}_{reg}$ is the matched-box L_1 loss; and GIoU_{3D} measures overlap in the joint temporal volume formed by adjacent frames. The three weights are $(2, 5, 2)$. Full matching and GIoU definitions are in the supplementary material.

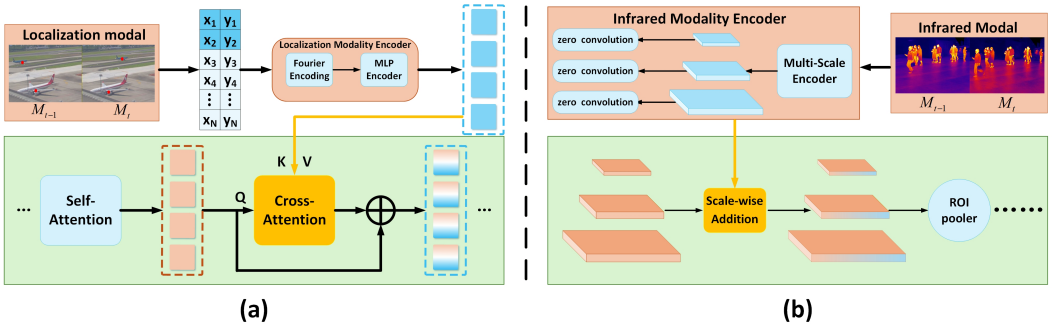


Figure 3: Encoding and injection of the two modality conditions. (a) Satellite positioning coordinates are mapped to sparse condition tokens by Fourier encoding and an MLP. As keys and values, they interact with proposal query features through cross-attention, then update proposal features by a residual connection. (b) A multiscale encoder extracts dense infrared features. Zero-initialized convolutions turn each scale into a residual added to the corresponding RGB feature pyramid, after which an ROI pooler extracts condition-aware regional features.

3.2 Multimodal Condition Injection

Auxiliary modalities enter the shared paired-box estimator through interfaces suited to their data forms, as shown in Fig. 3. GNSS positions provide sparse object-level spatial constraints, whereas infrared images provide dense thermal features. They are injected into proposal features and the RGB feature pyramid, respectively.

GNSS positioning condition. For frame $\tau \in \{t-1, t\}$, each GNSS entry contains a normalized coordinate projected onto the image plane, $p_\tau^j = (x_\tau^j, y_\tau^j)$, a positioning identity id_τ^j, and a validity flag m_τ^j. Entries are padded to a common length across frames; invalid and padded positions are excluded from cross-attention. Valid coordinates are represented by a multifrequency Fourier map:

$$\gamma(p_\tau^j) = \left[\sin(\omega_k x_\tau^j), \cos(\omega_k x_\tau^j), \sin(\omega_k y_\tau^j), \cos(\omega_k y_\tau^j)\right]_{k=0}^{L-1}, \tag{11}$$

where the frequencies ω_k are logarithmically spaced. An MLP maps the positional encoding to condition tokens, which form a framewise sequence:

$$c_\tau^j = \phi_{\mathrm{loc}}\big(\gamma(p_\tau^j)\big), \qquad C_\tau^{loc} = [c_\tau^1, \ldots, c_\tau^{N_{\mathrm{loc}}}]. \tag{12}$$

Condition encoding uses only positioning coordinates; identity labels are reserved for auxiliary association at inference.

In each Basic Block, ROI features extracted at the current boxes pass through proposal self-attention to yield query features $Q_\tau^{(r)}$. Cross-attention then uses the positioning tokens as keys and values and updates the proposal features residually:

$$\widetilde{Q}_\tau^{(r)} = Q_\tau^{(r)} + \mathrm{CrossAttn}\Big(Q_\tau^{(r)}, C_\tau^{loc}, C_\tau^{loc}\Big). \tag{13}$$

The updated features enter cross-frame fusion and the prediction heads. Repeating this operation in every Basic Block lets positioning continuously affect paired-box estimation. Coordinates may also initialize candidates, while positioning identities may assist trajectory matching; these inference-time functions are detailed in the supplementary material.

Infrared condition. An RGB backbone and FPN produce $P_\tau^{3:5}$, and a multiscale encoder processes the synchronized infrared image into aligned features $R_\tau^{3:5}$. To preserve their spatial structure, a zero-initialized 1×1 convolution injects the infrared features as residuals at each level $\ell \in \{3, 4, 5\}$ before ROI extraction:

$$\widehat{P}_\tau^\ell = P_\tau^\ell + g_\ell(R_\tau^\ell), \qquad \ell \in \{3, 4, 5\}. \tag{14}$$

Zero initialization gives $\widehat{P}_\tau^\ell = P_\tau^\ell$ at the start of training; the network then learns scale-specific infrared corrections. The paired-box estimator extracts ROI features from the fused pyramids in both frames and follows the same proposal interaction, cross-frame fusion, prediction, and sampling procedures as in the GNSS setting. Infrared ROI features can also assist later trajectory matching (supplementary material).

3.3 Modality-Specific Energy Guidance

Section 3.1 defines a unified energy-gradient update. We retain its paired-box matrix $Z \in \mathbb{R}^{N\times 8}$, whose ith row $z_i = [b_{t-1,i}, b_{t,i}]$ represents one candidate pair. Let $M_{\mathrm{loc}} = (O_{t-1}^{loc}, O_t^{loc})$ and $M_{\mathrm{ir}} = (I_{t-1}^{\mathrm{ir}}, I_t^{\mathrm{ir}})$ denote two-frame positioning and infrared information. Their energies measure spatial agreement with positioning data and cross-frame agreement of infrared regional features, respectively.

GNSS positioning energy: framewise independent guidance. From Section 3.2, let $\mathcal{P}_\tau = \{p_\tau^j \mid m_\tau^j = 1\}$ contain valid positions in frame τ, and let $c(b_{\tau,i})$ be the normalized center of a predicted box. We define

$$\begin{aligned} E_{\mathrm{loc}}(Z, M_{\mathrm{loc}}) &= \frac{1}{2N} \sum_{i=1}^{N} \sum_{\tau\in\{t-1,t\}} d_\tau(b_{\tau,i}), \\ d_\tau(b) &= \begin{cases} \min\limits_{p\in\mathcal{P}_\tau} \|c(b) - p\|_2^2, & \mathcal{P}_\tau \neq \varnothing, \\ 0, & \mathcal{P}_\tau = \varnothing. \end{cases} \end{aligned} \tag{15}$$

This energy measures the discrepancy between each box center and the nearest valid position in the same frame. The two framewise terms yield independent positional corrections for $b_{t-1,i}$ and $b_{t,i}$. Positioning identities do not enter the energy, although they can support the auxiliary association described in Section 3.2.

Infrared energy: cross-frame coupled guidance. Infrared images do not directly supply object coordinates; instead, box coordinates determine which image regions are sampled. Let

$$h_{\tau,i} = \mathcal{R}\big(R_\tau^{3:5}, b_{\tau,i}\big)\,, \tag{16}$$

where $R_\tau^{3:5}$ is the infrared feature pyramid of Section 3.2. The operator $\mathcal{R}$ performs differentiable bilinear regional sampling and feature aggregation according to the box coordinates. Moving or resizing a box thus changes the sampling locations and regional features. Define $\cos(u, v) = u^\top v/(\|u\|_2\|v\|_2 + \epsilon)$, with $\epsilon > 0$ for numerical stability. Given the existing correspondence of a candidate pair, the cross-frame infrared energy is

$$E_{\mathrm{ir}}(Z, M_{\mathrm{ir}}) = \frac{1}{N} \sum_{i=1}^{N} \left[1 - \cos(h_{t-1,i}, h_{t,i})\right]. \tag{17}$$

Each term depends on both regional features and therefore provides gradients for both boxes in the pair. The boxes are jointly constrained by feature consistency, but their correction directions and magnitudes need not be identical. Regional features reuse the existing infrared encoder, without a separately trained energy network.

At inference, we select $E_m = E_{\mathrm{loc}}$ or $E_m = E_{\mathrm{ir}}$ according to the input modality and substitute it into Eq. (7). The same gradient-update rule thus uses framewise positioning constraints or cross-frame infrared feature consistency. Corrections affect only intermediate sampling steps; the final network call again jointly predicts boxes, classes, and association scores before paired-box filtering and trajectory management.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and conditions. We evaluate GNSS-conditioned tracking on AGVS-T24 (Zhang et al., 2026). As synchronized GNSS data are unavailable, we construct perturbed two-dimensional positions from object annotations; trajectory identities support auxiliary association, while condition encoding uses only coordinates.

For infrared tracking, we use the pedestrian subset VT-Person-MOT of VT-MOT (Zhu et al., 2025), with visible images as primary input and infrared images as dense conditions. Dataset selection, splits, and preprocessing are detailed in the supplementary material.

Implementation details. We use YOLOX-X as the visual backbone with input resolution 896×1600. The backbone is frozen during training; only the paired-box estimator and multimodal condition branches are optimized. The estimator has six cascaded Basic Blocks; GNSS and infrared experiments use 1000 and 500 proposals, respectively. We train for 20 epochs with AdamW, an initial learning rate of 2.5×10^{-5}, weight decay 1×10^{-4}, and batch size 1. Mosaic augmentation is disabled for the final five epochs. The classification Focal Loss, box L_1 loss, and GIoU loss have weights 2, 5, and 2. Modality-specific hyperparameters are in the supplementary material.

4.2 COMPARISON WITH STATE-OF-THE-ART METHODS

Table 1: Comparison with existing multi-object tracking methods under different multimodal settings.

	GNSS-guided						Infrared-guided					
Methods	MOTA↑	IDF1↑	HOTA↑	AssA↑	DetA↑	IDs↓	MOTA↑	IDF1↑	HOTA↑	AssA↑	DetA↑	IDs↓
SORT	76.4	66.1	61.1	59.4	65.8	38	16.0	25.5	25.7	46.0	14.8	**81**
OC-SORT	76.4	66.1	61.0	59.3	65.7	37	16.0	24.5	24.6	42.1	14.7	88
BoT-SORT	78.8	66.0	61.5	58.8	**67.6**	45	19.7	29.9	30.0	48.9	19.0	113
Hybrid-SORT	76.4	65.5	60.3	57.8	65.7	39	15.9	24.3	24.4	41.5	14.7	93
ByteTrack	78.8	66.1	61.4	58.9	67.4	**33**	19.7	30.3	30.3	**50.0**	19.0	91
FairMOT	78.8	63.4	54.1	47.9	62.2	86	15.4	25.6	25.4	43.1	15.5	160
TrackFormer	64.1	56.0	54.4	53.1	60.3	132	19.8	27.0	26.6	35.8	20.3	205
TransTrack	73.6	61.1	56.4	53.9	63.6	339	17.2	28.1	27.4	40.9	18.9	219
DiffusionTrack	74.6	48.2	44.1	32.3	61.9	2806	27.0	27.7	25.6	26.4	25.7	1760
CDMMT	**82.4**	**75.5**	**61.6**	**59.7**	65.2	598	**52.8**	**48.5**	**41.8**	38.6	**46.0**	2132

GNSS-guided tracking. On AGVS-T24 (Table 1), CDMMT leads in MOTA, IDF1, HOTA, and AssA by 3.6, 9.4, 0.1, and 0.3 points over the strongest respective baselines. Against RGB-only DiffusionTrack, it raises MOTA from 74.6% to 82.4% and IDF1 from 48.2% to 75.5%, while reducing IDs from 2806 to 598. Its DetA of 65.2% remains below BoT-SORT's 67.6%, and its IDs do not lead, so the advantage is strongest in overall tracking and identity accuracy rather than every component metric. This full-system comparison includes positioning identities in auxiliary association; the ablation below isolates coordinate guidance.

Infrared-guided tracking. On VT-Person-MOT, CDMMT leads in MOTA, IDF1, HOTA, and DetA by 25.8, 18.2, 11.5, and 20.3 points over the strongest respective baselines. Relative to DiffusionTrack, MOTA rises from 27.0% to 52.8% and IDF1 from 27.7% to 48.5%; HOTA and DetA also rise from 25.6% and 25.7% to 41.8% and 46.0%. The larger DetA suggests that thermal information helps recover objects missed with RGB alone. AssA remains below ByteTrack's, and IDs are high. Broader detection coverage may increase identity-switch opportunities, although aggregate metrics cannot establish the cause; IDF1, AssA, and IDs should be read together.

4.3 QUALITATIVE ANALYSIS

Figure 4 shows that sparse GNSS positions support box localization and identity preservation, while dense infrared responses help recover and maintain trajectories. Missed detections and localization errors remain. Further examples are in the supplementary material.



Figure 4: Qualitative tracking results under GNSS positioning and infrared image conditions. The top row shows GNSS-conditioned tracking on AGVS-T24; the bottom row shows infrared-conditioned tracking on VT-Person-MOT. Box colors indicate different tracking identities.

Table 2: Ablation study on GNSS position condition quality.

Setting	MOTA↑	IDF1↑	HOTA↑	AssA↑	DetA↑	IDs↓
w/o condition	58.4	52.0	44.9	42.6	48.1	1068
w/ invalid condition	68.9	35.2	35.6	23.0	57.4	2680
w/ valid condition	**81.3**	**58.9**	**52.5**	**43.5**	**65.1**	**1051**

Table 3: Ablation study on infrared condition quality.

Setting	MOTA↑	IDF1↑	HOTA↑	AssA↑	DetA↑	IDs↓
w/o condition	11.5	18.7	20.5	33.5	12.9	785
w/ invalid condition	12.6	18.8	20.4	31.1	13.9	873
w/ valid condition	**52.8**	**48.5**	**41.8**	**38.6**	**46.0**	2132

4.4 Ablation Studies

We compare no condition, invalid condition (shuffled GNSS coordinates or all-zero infrared images), and valid condition (constructed GNSS positions or original infrared images). This separates useful modality information from the presence of the condition branch.

GNSS position condition ablation. With auxiliary tracking modules disabled, valid coordinates raise MOTA from 58.4 to 81.3, IDF1 from 52.0 to 58.9, and HOTA from 44.9 to 52.5 (Table 2). Shuffled coordinates lower IDF1 to 35.2 and raise IDs to 2680. Thus, the gain reflects spatial coordinates, not positioning identity numbers in the denoising network.

Infrared condition ablation. Real infrared images raise MOTA from 11.5% to 52.8%, IDF1 from 18.7% to 48.5%, and HOTA from 20.5% to 41.8% (Table 3); IDs also rise from 785 to 2132. All-zero infrared input resembles no condition: MOTA is 12.6% versus 11.5%, and IDF1 is 18.8% versus 18.7%. The large gap to valid infrared input links the gain to thermal information rather than merely enabling the branch. The aggregate metrics do not explain the higher IDs.

Across both modalities, the invalid controls show why condition quality matters. Shuffled GNSS positions improve MOTA over no condition (68.9 versus 58.4) but sharply reduce IDF1 (35.2 versus 52.0), so a detection gain alone does not imply reliable association. In contrast, zero infrared input changes MOTA by only 1.1 points and IDF1 by 0.1 points. Valid conditions improve both MOTA and IDF1 in each setting, supporting the use of modality information that remains aligned with the tracked objects.

5 Conclusion

We proposed CDMMT, a conditional diffusion framework for multimodal multi-object tracking. Modality-adapted condition encoding and injection incorporate heterogeneous auxiliary information into a shared paired-box denoising process for object localization and adjacent-frame association. In addition to feature-level condition injection, energy-guided sampling measures agreement between predicted boxes and modality information with differentiable conditional energies. Their gradients correct intermediate predictions, allowing modality constraints to influence later predictions through sampling-state updates. GNSS positioning and infrared images illustrate the encoding, injection, and guidance of sparse positional and dense image conditions; a depth-geometric extension is provided in the supplementary material. Experiments show that CDMMT effectively uses different forms of auxiliary information and improves overall tracking performance.

AI USE STATEMENT

AI tools assisted literature review, language editing, and Chinese–English translation.

REFERENCES

Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *Proceedings of the IEEE International Conference on Image Processing*, 2016.

Jiarui Cai, Mingze Xu, Wei Li, Yushi Xiong, Wei Xia, Zhuowen Tu, and Stefano Soatto. MeMOT: Multi-object tracking with memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8090–8100, 2022.

Jinkun Cao, Xinshuo Weng, Rawal Khirodkar, Kris Kitani Pang, and Kris Kitani. Observation-centric SORT: Rethinking SORT for robust multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Ruopeng Gao and Limin Wang. MeMOTR: Long-term memory-augmented transformer for multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9901–9910, 2023.

Ruopeng Gao, Jinqing Qi, and Limin Wang. Multiple object tracking as ID prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27883–27893, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Yaxuan Hu, Jie Hua, Zhen Han, Hua Zou, Gang Wu, and Zhongyuan Wang. DiffusionMOT: A diffusion-based multiple object tracker. *IEEE Transactions on Neural Networks and Learning Systems*, 36(10):18203–18217, 2025.

Cong Huang, Shoudong Han, Minglei He, Wei-Shi Zheng, and Yunchao Wei. DeconfuseTrack: Dealing with confusion for multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19290–19299, 2024.

Jie Jiang, Zhe Wang, Ming Zhao, Yue Li, and Dong Jiang. SAM2MOT: A novel paradigm of multi-object tracking by segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5388–5396, 2026a.

Lin Jiang, Zhe Wang, Shiming Yin, Guoqing Ma, Peng Zhang, and Bin Wu. ConsistencyTrack: A robust multi-object tracker with a generation strategy of consistency model. *Pattern Recognition*, 171:112225, 2026b.

Chenglong Li, Andong Lu, Aihua Zheng, Zhipeng Tu, and Jin Tang. Multi-adapter RGBT tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 2262–2270, 2019.

Hao Liu, Cheng Zhang, Bin Fan, and Jun Xu. Pro2Diff: Proposal propagation for multi-object tracking via the diffusion model. *IEEE Transactions on Image Processing*, 33:6508–6520, 2024.

Zhichao Liu, Xiaotong Wang, Chao Wang, Wenyu Liu, and Xiang Bai. SparseTrack: Multi-object tracking by performing scene decomposition based on pseudo-depth. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(5):4870–4882, 2025.

Run Luo, Zikai Song, Lintao Ma, Jinlin Wei, Wei Yang, and Min Yang. DiffusionTrack: Diffusion model for multi-object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3991–3999, 2024.

Yiming Luo, Xiaojie Guo, Ming Dong, and Jun Yu. Learning modality complementary features with mixed attention mechanism for RGB-T tracking. *Sensors*, 23(14):6609, 2023.

Weiyi Lv, Yuhang Huang, Ning Zhang, Ruei-Sung Lin, Mei Han, and Dan Zeng. DiffMOT: A real-time diffusion-based multiple object tracker with non-linear prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19321–19330, 2024.

Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixé, and Christoph Feichtenhofer. TrackFormer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8844–8854, 2022.

Ramin Nabati and Hairong Qi. CenterFusion: Center-based radar and camera fusion for 3d object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1527–1536, 2021.

Ramin Nabati, Luke Harris, and Hairong Qi. CFTrack: Center-based radar and camera fusion for 3d multi-object tracking. In *Proceedings of the IEEE Intelligent Vehicles Symposium Workshops*, pp. 243–248, 2021.

Zheng Qin, Sanping Zhou, Le Wang, Jinghai Duan, Gang Hua, and Wei Tang. MotionTrack: Learning robust short-term and long-term motions for multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17939–17948, 2023.

Kha Gia Quach, Phong Nguyen, Chi Nhan Duong, Tien D. Bui, and Khoa Luu. Depth perspective-aware multiple object tracking. In *Engineering Applications of AI and Swarm Intelligence*, pp. 181–205. Springer, 2024.

Hao Ren, Shoudong Han, Huilin Ding, Ziwen Zhang, Hong Wang, and Fa Wang. Focus on details: Online multi-object tracking with diverse fine-grained representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11289–11298, 2023.

Jenny Seidenschwarz, Guillem Brasó, Víctor Castro Serrano, Ismail Elezi, and Laura Leal-Taixé. Simple cues lead to a strong multi-object tracker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13813–13823, 2023.

Kyujin Shim, Kwangjin Ko, Youngbae Yang, and Changick Kim. Focusing on tracks for online multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11687–11696, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proceedings of the International Conference on Learning Representations*, 2021.

Xiao Wang, Xiujun Shu, Shiliang Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng Wu. MFGNet: Dynamic modality-aware filter generation for RGB-T tracking. *IEEE Transactions on Multimedia*, 25:4335–4348, 2023.

Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *Proceedings of the IEEE International Conference on Image Processing*, 2017.

Qing Xu, Lei Wang, Weiguo Sheng, Yiming Wang, Chao Xiao, Chenguang Ma, and Wei An. Heterogeneous graph transformer for multiple tiny object tracking in RGB-T videos. *IEEE Transactions on Multimedia*, 26:9383–9397, 2024.

Mingzhan Yang, Guang Han, Bin Yan, Wenhua Zhang, Jinqing Qi, Huchuan Lu, and Dong Wang. Hybrid-SORT: Weak cues matter for online multi-object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6504–6512, 2024.

Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. MOTR: End-to-end multiple-object tracking with transformer. In *Proceedings of the European Conference on Computer Vision*, pp. 659–675, 2022.

Xinyu Zhang, Tao Li, Zhen Tang, Gang Wang, and Rui Huang. A multiple aircraft tracking dataset for airport traffic surveillance. *IEEE Transactions on Intelligent Transportation Systems*, 27(4): 4017–4030, 2026. doi: 10.1109/TITS.2026.3658559.

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 *Proceedings of the European Conference on Computer Vision*, 2022.

Yuang Zhang, Tiancai Wang, and Xiangyu Zhang. MOTRv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22056–22065, 2023.

Xingyi Zhou, Tianwei Yin, Vladlen Koltun, and Philipp Krähenbühl. Global tracking transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8771–8780, 2022.

Yabin Zhu, Qiang Wang, Chenglong Li, Jin Tang, Chuang Gu, and Zhi Huang. Visible-thermal multiple object tracking: Large-scale video dataset and progressive fusion approach. *Pattern Recognition*, 161:111330, 2025. doi: 10.1016/j.patcog.2024.111330.