SPECHYPER-TRACK: RELATION-SPECIFIC SPECTRAL HYPERGRAPH REASONING FOR ONLINE MULTI-OBJECT TRACKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Relational reasoning methods often assume that all interactions share identical propagation behaviors, while real-world systems contain relations with different semantic and spectral characteristics. Online multi-object tracking (MOT) provides a challenging instance of this problem, where cooperative motion patterns and identity-conflicting interactions coexist. We introduce SpecHyper-Track, a relation-specific spectral hypergraph framework with stage-aware frequency routing for identity-preserving association. The framework constructs semantic hypergraphs with homogeneous and heterogeneous hyperedges to model cooperative and competitive relations, respectively, enabling relation-dependent spectral propagation. Cooperative relations are enhanced through low-frequency smoothing, whereas competitive relations preserve identity-sensitive variations through high-frequency responses. Leveraging tight-framelet decomposition, we derive complementary frequency-aware representations adaptively dispatched based on trajectory status: high-frequency features improve discrimination for established trajectories, while fused low- and high-frequency features provide stable context for newly initialized trajectories. The proposed framework couples semantic relational modeling, spectral representation learning, and state-aware association in a unified manner. Evaluations on MOT16, MOT17 and DanceTrack yield substantial improvements on association-oriented metrics including HOTA, IDF1 and AssA in crowded interactive scenarios. Code is available at `https://anonymous.4open.science/r/SpecHyper-Track-1869/`.

1 INTRODUCTION

Relational reasoning provides a natural way to model interactions among multiple entities. Graph neural networks capture structured dependencies through message passing Kipf & Welling (2017), but typically apply shared propagation to connected nodes, limiting their ability to represent relation-specific semantics Hu et al. (2020). This issue is particularly relevant to online multi-object tracking (MOT), where coordinated targets provide cooperative motion context Pellegrini et al. (2009), while spatially adjacent or visually similar targets introduce competitive interactions requiring identity preservation Milan et al. (2013). Most modern MOT systems follow the tracking-by-detection paradigm and associate detections with trajectories using motion, geometry, appearance, or confidence cues Luo et al. (2021). Nevertheless, ambiguous association remains a major source of identity switches in crowded scenes Huang et al. (2024b).

Existing trackers improve association through motion modeling Cao et al. (2023), discriminative appearance representations Wojke et al. (2017), confidence-aware matching Zhang et al. (2022), or transformer-based trajectory modeling Meinhardt et al. (2022). However, many methods still formulate association primarily through individual trajectory–detection affinities Brasó & Leal-Taixé (2020). Graph matching introduces structural context He et al. (2021), and global attention enables information exchange across objects and frames Zhou et al. (2022), yet their interactions are predominantly represented through pairwise relations. Such formulations may overlook group-wise dependencies shared by multiple targets, which hypergraphs can naturally encode through variable-order hyperedges Wen et al. (2019).

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.

High-order connectivity alone, however, does not resolve the semantic heterogeneity of target interactions. Conventional graph convolution is largely homophily-oriented Zhu et al. (2020), while hypergraphs extend pairwise edges to multi-node relations Feng et al. (2019) but commonly process different hyperedges using shared aggregation operators Huang & Yang (2021). Consequently, cooperative and competitive interactions may receive similar propagation despite requiring different representations. Relation-aware learning suggests that distinct relation types should employ dedicated transformations Yu et al. (2023): cooperative relations should promote contextual consistency, whereas competitive relations should preserve identity-sensitive differences.

This semantic distinction also admits a spectral interpretation. Homophilic interactions favor low-frequency smoothing, whereas heterophilic relations benefit from high-frequency responses that preserve feature variations Luan et al. (2022). Adaptive spectral filtering further supports relation-dependent frequency processing Zhang et al. (2025). Recent hypergraph methods exploit complementary low- and high-frequency components through tight framelets Li et al. (2025a) and address homophily-biased propagation Li et al. (2026b), but frequency processing is generally not conditioned jointly on relation semantics and trajectory states. This is restrictive for MOT because locally varying evidence benefits from node-adaptive filtering Zheng et al. (2024), while trajectories at different lifecycle stages possess different levels of historical support Pang et al. (2023).

To address these limitations, we propose a relation-specific spectral hypergraph framework with stage-aware frequency routing for online MOT. Homogeneous hyperedges use motion-spatial consistency as a cue for contextual cooperation, while heterogeneous hyperedges use cross-group appearance ambiguity as a cue for potential competition. The two relation types are associated with complementary low- and high-frequency responses, followed by tight-framelet decomposition. The resulting representations are further routed according to trajectory states: high-frequency features strengthen discrimination for established trajectories, whereas fused low/high-frequency features provide more stable context for newly initialized trajectories. The main contributions are summarized as follows:

- We formulate relation-specific spectral hypergraph reasoning for MOT by separating cooperative and competitive high-order interactions and assigning them complementary low- and high-frequency propagation for contextual smoothing and identity discrimination.
- We introduce stage-aware frequency routing that couples spectral representations with trajectory states, using high-frequency features for established trajectories and fused representations for newly initialized trajectories.
- We develop a training-free spectral refinement module that integrates into existing online association pipelines without additional training or changes to trajectory management.

2 Proposed Tracking Method

2.1 Problem Definition and Overall Framework

We propose SpecHyper-Track, a relation-specific spectral hypergraph framework for online multi-object tracking. At frame t, we formulate data association as a relation-aware representation refinement problem. Rather than relying solely on pairwise trajectory–detection affinity, we first construct a unified relational structure over current observations, temporal context, and motion predictions, and then refine node representations through relation-specific spectral reasoning.

Let $\mathcal{D}^t$, $\mathcal{C}^{t-1}$, and $\mathcal{T}^t$ denote the current detection nodes, previous-frame contextual nodes, and Kalman-predicted trajectory nodes, respectively. The unified node set is defined as $\mathcal{V}^t = \mathcal{D}^t \cup \mathcal{C}^{t-1} \cup \mathcal{T}^t = \{n_i\}_{i=1}^{N}$, where n_i denotes the i-th node and $N = |\mathcal{V}^t|$. Current detections, previous-frame context, and Kalman-predicted trajectories are represented as distinct nodes according to their roles, even when they correspond to the same target. Each node n_i is associated with appearance, spatial, and motion information. Let $\hat{f}_i \in \mathbb{R}^d$ denote the normalized appearance feature of n_i. The node features are stacked as $X = \left[\hat{f}_1; \hat{f}_2; \cdots; \hat{f}_N\right] \in \mathbb{R}^{N \times d}$, where $N = |\mathcal{V}^t|$ and d denotes the feature dimension.

Our key premise is that high-order interactions in crowded scenes carry different semantics and therefore should not share a uniform propagation mechanism. Motion-consistent targets provide cooperative contextual evidence, whereas potentially confusable targets across different motion groups

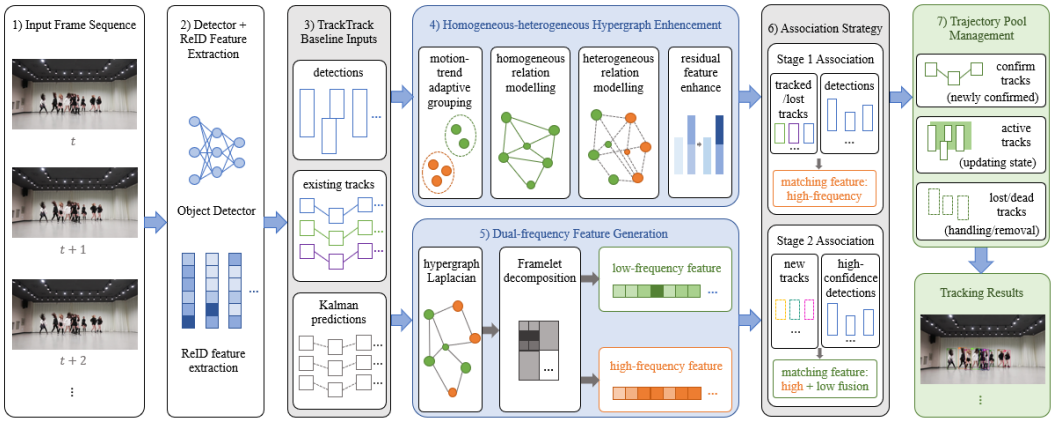


Figure 1: Overall framework of the proposed SpecHyper-Track. Current detections, previous-frame contextual nodes, and Kalman-predicted trajectories form a unified node set. Motion-trend grouping and relation-specific propagation generate enhanced representations, which are decomposed into complementary frequency components. High-frequency and fused representations are respectively routed to the two association stages, followed by trajectory management.

form competitive interactions that require stronger identity separation. We therefore construct two relation-specific hypergraphs, $\mathcal{G}^h = (\mathcal{V}^t, \mathcal{E}^h)$ and $\mathcal{G}^g = (\mathcal{V}^t, \mathcal{E}^g)$, to model cooperative intra-group relations and competitive inter-group relations, respectively.

As illustrated in Fig. 1, SpecHyper-Track consists of three components: semantic relation modeling, relation-specific spectral enhancement, and stage-aware frequency routing. Within a track-oriented multi-stage association pipeline Shim et al. (2025), we introduce relation-specific hypergraph reasoning and stage-aware spectral representations for appearance association. Motion prediction, assignment, and track management remain unchanged. After matching, trajectory memory is updated with the relation-enhanced feature before frequency decomposition.

2.2 Semantic Relation Discovery

The goal of semantic relation discovery is to identify motion-consistent nodes for constructing cooperative high-order relations. We jointly consider motion direction and spatial compatibility, since motion-consistent nearby targets can provide reliable contextual evidence in crowded scenes. For two distinct nodes n_i and n_j, the motion-direction similarity is defined as

$$S_m(i,j) = \begin{cases} \frac{1}{2}\left(\frac{v_i^{\mathrm{T}} v_j}{\|v_i\|_2 \|v_j\|_2} + 1\right), & \min\{\|v_i\|_2, \|v_j\|_2\} > \epsilon, \\ 0.5, & \text{otherwise}, \end{cases} \tag{1}$$

where v_i and v_j denote the node velocities. The cosine similarity is mapped to $[0, 1]$, while 0.5 is used when motion is unreliable. Spatial compatibility is measured by

$$S_s(i,j) = \exp\left(-\frac{\|c_i - c_j\|_2}{\sigma}\right), \tag{2}$$

where c_i and c_j denote the center positions of nodes n_i and n_j, respectively. Combining motion consistency, spatial proximity, and distance suppression, the grouping score is

$$G(i,j) = \omega_m S_m(i,j) + \omega_s S_s(i,j) - \omega_d \min\left\{\frac{\|c_i - c_j\|_2}{r_{\max}}, 1\right\} \tag{3}$$

where ω_m, ω_s, and ω_d control the contributions of the three terms, and $r_{\max}$ denotes the maximum spatial distance considered for motion grouping. A larger $G(i,j)$ indicates stronger motion-spatial compatibility. To avoid chain-like grouping, node n_i is assigned to an existing motion group $\mathcal{G}_k$ only if $|\mathcal{G}_k| < K_{\max}$ and it is compatible with every group member:

$$n_i \to \mathcal{G}_k \iff \begin{cases} G(i,j) \geq \theta_g, \\ \|c_i - c_j\|_2 \leq r_{\max}, \end{cases} \quad \forall n_j \in \mathcal{G}_k. \tag{4}$$

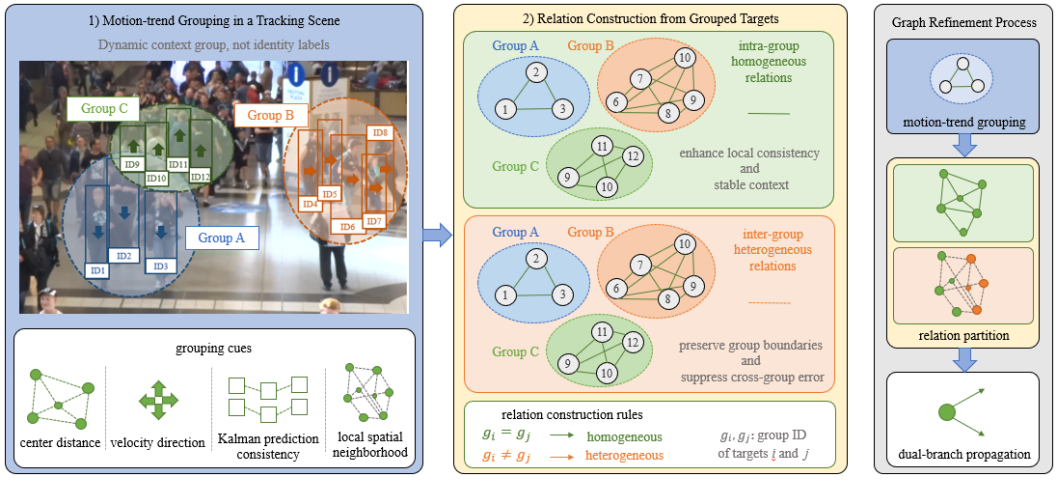


Figure 2: Construction of homogeneous/heterogeneous hyperedges and graph refinement. Motion-consistent nodes form homogeneous hyperedges, while ambiguous nodes from neighboring groups form heterogeneous hyperedges. The resulting relations are refined through dual-branch spectral propagation.

Unlike single-link grouping, which may connect incompatible nodes through intermediate targets, Eq. (4) enforces complete-link consistency by requiring each newly added node to be compatible with all existing group members. Consequently, each resulting group satisfies

$$\min_{\substack{n_i,n_j\in\mathcal{G}_k \\ i\neq j}} G(i,j) \geq \theta_g, \qquad \max_{\substack{n_i,n_j\in\mathcal{G}_k \\ i\neq j}} \|c_i - c_j\|_2 \leq r_{\max}. \tag{5}$$

This property prevents chain-like connections between distant or motion-inconsistent nodes, providing more reliable cooperative relations for subsequent hypergraph construction.

2.3 Semantic Hypergraph Construction

Based on the discovered motion-consistent groups, we construct two types of semantic hyperedges. Homogeneous hyperedges encode within-group motion-spatial consistency, whereas heterogeneous hyperedges encode cross-group appearance ambiguity. These serve as proxies for contextual cooperation and potential competition, rather than ground-truth identity relations. The overall construction is illustrated in Fig. 2.

Let $\mathcal{E}^h$ and $\mathcal{E}^g$ denote the homogeneous and heterogeneous hyperedge sets, respectively. For each relation type $r \in \{h, g\}$, the corresponding incidence matrix is defined as $H^r_{ij} = 1$ when $n_i \in e^r_j$, and 0 otherwise, where e^r_j denotes the j-th hyperedge of relation type r. Each column of H^r corresponds to one variable-order relation, while the two incidence matrices preserve the distinction between cooperative and competitive interactions.

For each motion group $\mathcal{G}_j$ with $N_j = |\mathcal{G}_j| \geq 2$, we construct its homogeneous hyperedge e^h_j. Its weight is defined by the mean pairwise compatibility:

$$w^h_j = \max\left\{ \frac{2}{N_j(N_j-1)} \sum_{\substack{n_a,n_b\in\mathcal{G}_j \\ a<b}} G(a,b), \epsilon_w \right\}, \tag{6}$$

where $G(a,b)$ is defined in Eq. (3), and $\epsilon_w > 0$ prevents degenerate weights. Heterogeneous hyperedges are constructed to model ambiguous interactions across distinct motion groups. For normalized appearance features, we define the appearance similarity as

$$S_a(i,j) = \mathrm{clip}_{[0,1]}((\hat{f}_i^{\mathrm{T}} \hat{f}_j + 1)/2). \tag{7}$$

For two distinct motion groups $\mathcal{G}_p$ and $\mathcal{G}_q$, the eligible cross-group pairs are defined as

$$\mathcal{Q}_{pq} = \{(i,j) \,|\, n_i \in \mathcal{G}_p,\ n_j \in \mathcal{G}_q,\ d_{ij} \le r_h,\ (S_a(i,j) \ge \theta_a \ \vee \ d_{ij} \le r_c)\}. \tag{8}$$

where $d_{ij} = \|c_i - c_j\|_2$, r_h denotes the cross-group search radius, r_c is the close-distance threshold, and θ_a controls appearance ambiguity. This criterion retains nearby cross-group pairs that are either appearance-confusable or sufficiently close. For each eligible pair, its confusion degree is defined as

$$C(i,j) = \lambda_a S_a(i,j) + \lambda_s S_s(i,j), \tag{9}$$

where $S_s(i,j)$ is defined in Eq. (2), and λ_a and λ_s control the contributions of appearance and spatial cues. The eligible pairs in $\mathcal{Q}_{pq}$ are ranked according to $C(i,j)$, and their endpoints are progressively merged into a heterogeneous hyperedge:

$$e^g_{pq} = \mathrm{GreedyUnion}_{K_{\max}}\left(\{\{n_i, n_j\} \,|\, (i,j) \in \mathcal{Q}_{pq}\}; C\right), \tag{10}$$

where $\mathrm{GreedyUnion}_{K_{\max}}$ accepts pairs in descending order of $C(i,j)$ while limiting the hyperedge order to $K_{\max}$. If $\mathcal{Q}_{pq} = \emptyset$, no heterogeneous hyperedge is created. Let $\mathcal{P}_{pq} \subseteq \mathcal{Q}_{pq}$ denote the accepted pairs used to construct e^g_{pq}. For non-empty $\mathcal{P}_{pq}$, the heterogeneous hyperedge weight is defined as

$$w^g_{pq} = \max\left\{ \frac{1}{|\mathcal{P}_{pq}|} \sum_{(i,j)\in\mathcal{P}_{pq}} C(i,j), \epsilon_w \right\}. \tag{11}$$

The resulting homogeneous and heterogeneous hypergraphs encode cooperative and competitive high-order relations separately, providing the structural basis for relation-specific spectral propagation.

2.4 Graph Refinement via Training-Free Dual-Relation Spectral Propagation

Homogeneous and heterogeneous hypergraphs encode different interaction semantics and therefore require different propagation responses. Homogeneous relations favor contextual consistency, whereas heterogeneous relations should preserve identity-sensitive differences.

We thus construct an independent spectral operator for each relation type $r \in \{h, g\}$. The node and hyperedge degree matrices are defined as $D^r_v(i,i) = \sum_{j=1}^{|\mathcal{E}^r|} H^r_{ij} w^r_j$ and $D^r_e(j,j) = \sum_{i=1}^{N} H^r_{ij}$, where H^r and w^r_j denote the incidence matrix and the j-th hyperedge weight of relation r, respectively. Following the normalized Zhou formulation, the relation-specific propagation operator is

$$\Theta_r = (D^r_v)^{-\frac{1}{2}} H^r W^r_e (D^r_e)^{-1} (H^r)^{\mathrm{T}} (D^r_v)^{-\frac{1}{2}}, \tag{12}$$

where $W^r_e = \mathrm{diag}(w^r_1, \ldots, w^r_{|\mathcal{E}^r|})$. For zero-degree nodes, the corresponding entries of $(D^r_v)^{-1/2}$ are set to zero. The normalized hypergraph Laplacian is $L_r = I - \Theta_r$. For nonnegative hyperedge weights, the normalized Zhou operator Θ_r is symmetric positive semidefinite with spectrum contained in $[0,1]$. Consequently, L_r is also positive semidefinite with eigenvalues in $[0,1]$. Let $u_{r,i}$ denote a shared eigenvector of L_r and Θ_r, satisfying $L_r u_{r,i} = \lambda_{r,i} u_{r,i}$. Thus, the following formula is obtained as $\Theta_r u_{r,i} = (1 - \lambda_{r,i}) u_{r,i}$.

Hence, Θ_r has the spectral response $h_\Theta(\lambda) = 1 - \lambda$, which favors low-frequency components, whereas L_r has $h_L(\lambda) = \lambda$, which emphasizes high-frequency variations. This complementary behavior naturally matches the two relation semantics: homogeneous relations benefit from smoothing, while heterogeneous relations require identity-preserving sharpening. Accordingly, the homogeneous branch uses $(\Theta_h - I)X = -L_h X$, whereas the heterogeneous branch uses $(I - \Theta_g)X = L_g X$. To restrict each correction to participating nodes, we define $M_r = \mathrm{diag}(m^r_1, \ldots, m^r_N)$, where $m^r_i = 1$ when $D^r_v(i,i) > 0$ and 0 otherwise. The dual-relation enhancement is formulated as

$$X_b = \mathrm{Norm}\Big(X + \alpha M_h \tanh\big((\Theta_h - I)X\big) \ + \beta M_g \tanh\big((I - \Theta_g)X\big)\Big), \tag{13}$$

where α and β are fixed, non-learnable coefficients, and $\mathrm{Norm}(\cdot)$ denotes row-wise ℓ_2 normalization. The element-wise $\tanh(\cdot)$ bounds the residual magnitude and reduces the influence of noisy relations.

Ignoring the nonlinear bounding and relation masks, the homogeneous and heterogeneous updates become $(I - \alpha L_h)X$ and $(I + \beta L_g)X$, with spectral responses $h_h(\lambda) = 1 - \alpha\lambda$ and $h_g(\lambda) = 1 + \beta\lambda$, respectively. Thus, the homogeneous branch suppresses high-frequency variations, whereas the heterogeneous branch amplifies them. The resulting relation-enhanced representation X_b is subsequently used as the input to relation-specific tight-framelet decomposition.

2.5 Relation-Specific Spectral Frequency Generation

The preceding dual-relation propagation provides coarse relation-aware spectral refinement. To further obtain explicit frequency-selective representations, we apply tight-framelet decomposition to the two semantic hypergraphs, extracting low-frequency context from homogeneous relations and high-frequency identity-sensitive information from heterogeneous relations.

For relation type $r \in \{h, g\}$, the normalized hypergraph Laplacian admits the spectral decomposition

$$L_r = U_r \Lambda_r U_r^{\mathrm{T}}, \qquad 0 \le \lambda_{r,1} \le \cdots \le \lambda_{r,N} \le 1, \tag{14}$$

where $\Lambda_r = \mathrm{diag}(\lambda_{r,1}, \ldots, \lambda_{r,N})$. Small eigenvalues correspond to smooth relational signals, whereas larger eigenvalues represent stronger variations across relational boundaries.

We adopt Haar tight-frame filters with low-pass and high-pass responses $\widehat{a}(\xi) = \cos(\xi/2)$ and $\widehat{b}(\xi) = \sin(\xi/2)$, where ξ denotes the scale-normalized spectral variable. Their squared responses satisfy $|\widehat{a}(\xi)|^2 + |\widehat{b}(\xi)|^2 = 1$, providing a complementary partition of spectral energy. Accordingly, the low-pass response emphasizes smooth components, whereas the high-pass response preserves rapid relational variations. This matches our relation semantics: cooperative relations favor low-frequency context, while competitive relations require high-frequency identity discrimination. For efficient filtering, the scale-dependent low-pass and high-pass responses are approximated by K_c-order Chebyshev polynomials:

$$p_{\phi,r,j}(L_r) = \frac{c_{\phi,0}^{(j)}}{2} I + \sum_{k=1}^{K_c - 1} c_{\phi,k}^{(j)} T_k(\widetilde{L}_{r,j}), \qquad \phi \in \{a, b\}, \tag{15}$$

where $T_k(\cdot)$ denotes the k-th order Chebyshev polynomial, $p_{a,r,j}$ and $p_{b,r,j}$ approximate the scale-dependent low-pass and high-pass responses, respectively, and $\widetilde{L}_{r,j}$ is the corresponding rescaled Laplacian on the Chebyshev approximation interval. This avoids explicit eigendecomposition during framelet filtering. Let $A_{r,0} = I$. The multilevel low-pass cascade and high-pass operators are recursively constructed as

$$W_{(1,j,r)} = p_{b,r,j}(L_r) A_{r,j-1} \text{ and } A_{r,j} = p_{a,r,j}(L_r) A_{r,j-1}, \quad j = 1, \cdots, J, \tag{16}$$

where $p_{a,r,j}$ and $p_{b,r,j}$ approximate the Haar low-pass and high-pass responses on scale j, respectively, and $W_{(0,J,r)} = A_{r,J}$. According to relation semantics, the homogeneous low-frequency component is reconstructed as

$$F_l = W_{(0,J,h)}^{\mathrm{T}} W_{(0,J,h)} X_b, \tag{17}$$

whereas the heterogeneous high-frequency component aggregates the high-pass bands:

$$F_h = \sum_{j=1}^{J} W_{(1,j,g)}^{\mathrm{T}} W_{(1,j,g)} X_b. \tag{18}$$

Thus, F_l captures cooperative contextual information, while F_h preserves identity-sensitive variations induced by competitive relations. Using the relation masks M_h and M_g introduced in Sec. 2.4, the reconstructed components are incorporated through bounded residual corrections:

$$\begin{cases} X_l = \mathrm{Norm}\left(X_b + \gamma_l M_h \tanh(F_l - X_b)\right), \\ X_h = \mathrm{Norm}\left(X_b + \gamma_h M_g \tanh(F_h)\right), \end{cases} \tag{19}$$

where γ_l and γ_h are fixed residual coefficients. Here, X_l and X_h denote the low-frequency-refined and high-frequency-enhanced representations, respectively. Both retain the base representation X_b through residual connections, rather than containing only a single frequency band. Finally, complementary representations are fused as

$$X_f = \mathrm{Norm}\left(\eta X_h + (1 - \eta) X_l\right), \tag{20}$$

where $\eta \in [0, 1]$ controls the high-frequency contribution. The resulting X_h and X_f are subsequently routed to different association stages according to trajectory states. Established trajectories use high-frequency-enhanced representations, whereas newly initialized trajectories use fused representations with additional contextual information. Accordingly,

$$(z_i^{(s)}, z_j^{(s)}) = \begin{cases} (x_{h,i}, x_{h,j}), & s = 1, \\ (x_{f,i}, x_{f,j}), & s = 2. \end{cases} \tag{21}$$

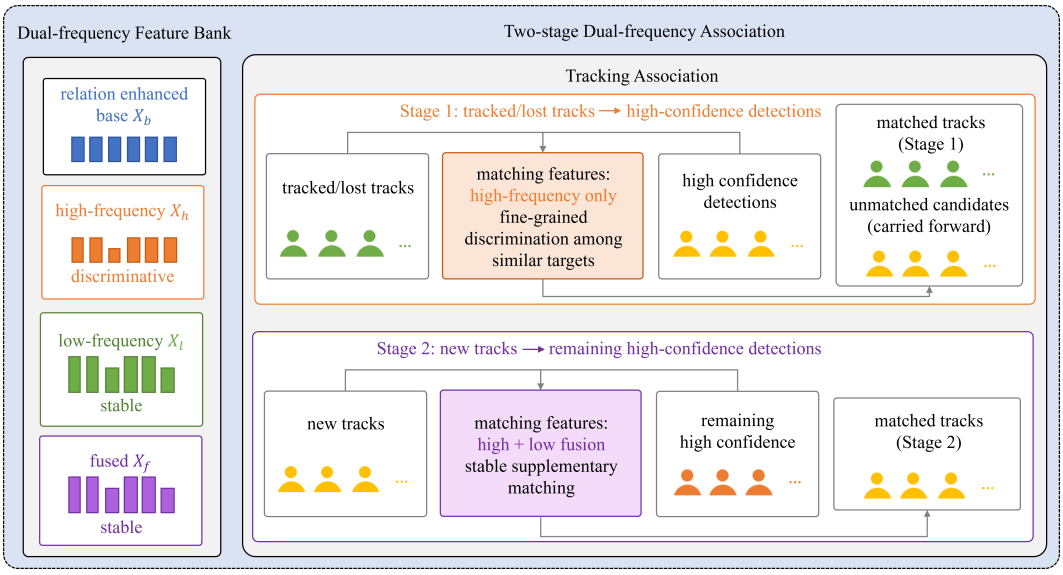


Figure 3: Dual-frequency routing and two-stage association. Stage 1 uses high-frequency-enhanced X_h for tracked/lost trajectories; Stage 2 uses fused X_f for new trajectories. "High-frequency only" means selecting X_h without fusion with X_l; both retain the base representation X_b. Trajectory memory is updated using X_b.

where $s \in \{1, 2\}$ denotes the association stage. The stage-specific appearance cost is computed using cosine distance:

$$d_{\text{app}}^{(s)}(i,j) = 1 - \frac{(z_i^{(s)})^\top z_j^{(s)}}{\|z_i^{(s)}\|_2 \|z_j^{(s)}\|_2}. \tag{22}$$

The resulting frequency-aware representations are incorporated into the stage-specific appearance cost, while the geometric, confidence, and motion terms are retained:

$$C_{ij}^{(s)} = \omega_{\text{iou}} d_{\text{HMIoU}}(T_i, d_j) + \omega_{\text{app}} d_{\text{app}}^{(s)}(i,j) + \omega_{\text{conf}} d_{\text{conf}}(T_i, d_j) + \omega_{\text{ang}} d_{\text{ang}}(T_i, d_j). \tag{23}$$

As illustrated in Fig. 3, Stage 1 uses X_h for tracked and lost trajectories to distinguish competing candidates, whereas Stage 2 uses X_f for new trajectories to supplement limited historical evidence with contextual information.

$$\bar{x}_i^t = \text{Norm}\left(\mu_j^t \bar{x}_i^{t-1} + (1 - \mu_j^t) x_{b,j}^t\right), \tag{24}$$

where $\bar{x}_i^t$ denotes the updated trajectory memory, $x_{b,j}^t$ is the relation-enhanced base feature of the matched detection, and μ_j^t is a confidence-adaptive memory coefficient. Newly initialized trajectories use the corresponding base feature $x_{b,j}^t$.

3 Experimental Study

We evaluate our method on MOT16, MOT17, and DanceTrack. MOT16 and MOT17 are standard pedestrian-tracking benchmarks containing crowded scenes, frequent occlusions, and camera motion. We use MOT17 as the primary benchmark for ablation studies. DanceTrack contains targets with highly similar appearances, non-linear motion, and frequent interactions, making it particularly suitable for evaluating identity association Sun et al. (2022).

We evaluate tracking performance using HOTA, DetA, AssA, MOTA, IDF1, and identity switches (IDs) with the TrackEval toolkit. Following common MOT evaluation practice, HOTA measures overall tracking quality, while AssA and IDF1 emphasize association and identity preservation Luiten et al. (2021). Since our method primarily targets ambiguous association, we focus on HOTA, AssA, IDF1, and IDs in the following analysis.

Table 1: Ablation study of homogeneous–heterogeneous relation modeling on the MOT17 validation set using GMTracker as the host tracker.

Setting	Hypergraph	Motion grouping	Relation update	HOTA↑	IDF1↑	MOTA↑	AssA↑	DetA↑
(a) Hypergraph only	✓	–	–	60.9	67.8	62.1	56.2	67.1
(b) Group-aware hypergraph	✓	✓	–	61.6	58.9	52.4	56.3	67.4
(c) Global relation replacement	✓	✓	Global	62.2	48.7	54.6	56.8	68.3
(d) Partial residual correction	✓	✓	Residual	**62.8**	**71.1**	**62.3**	**57.9**	**68.8**

Table 2: Ablation study of spectral feature construction and stage-aware association on the MOT17 validation set.

Method	HOTA↑	IDF1↑	MOTA↑	AssA↑	DetA↑	IDs↓
(a) Base tracker	66.0	81.1	77.9	68.1	64.5	1487
(b) Dual-relation enhancement	68.9	84.9	79.5	72.5	65.9	1376
(c) Analysis-domain framelet features	65.8	80.8	78.0	67.6	64.5	1469
(d) Vertex-domain framelet reconstruction	66.0	81.5	78.3	67.7	64.8	1546
(e) Stage-1 high-frequency routing	69.1	84.8	**79.9**	72.5	66.0	1198
(f) Dual-frequency cascade	**69.3**	**85.7**	79.8	73.1	66.1	1189
(g) Full model	**69.3**	**85.7**	79.8	**73.2**	**66.2**	**1101**

3.1 Ablation Studies

We conduct ablation studies on the MOT17 validation set to evaluate homogeneous–heterogeneous relation modeling and stage-aware dual-frequency association. Within each table, all variants use identical detections, ReID features, association thresholds, and evaluation protocols. Table 1 evaluates the relation-modeling module with GMTracker as the host tracker, whereas Table 2 evaluates the complete design with our principal tracking framework. Since the host trackers differ, absolute results across the two tables are not directly comparable.

Effect of homogeneous–heterogeneous relation modeling. We first investigate how motion grouping and different relation-update strategies affect tracking performance. The results in Table 1 show that explicitly introducing relation structure does not necessarily improve identity association when the relational information is injected too aggressively.

In particular, motion grouping provides useful structural context, but directly propagating or globally replacing the original appearance representation may lead to excessive feature mixing and degrade identity discrimination. This effect is especially evident for the global-replacement variant, which substantially reduces IDF1. In contrast, the partial-residual strategy preserves the original ReID representation while introducing only bounded relation-dependent corrections. It achieves the best performance, reaching 62.8 HOTA and 71.1 IDF1. Effective relation modeling therefore requires informative high-order relations and controlled integration into identity features.

Table 3: Sensitivity analysis of the relation weights α and β on the MOT17 validation set. Bold numbers indicate the best performance.

α	β	HOTA↑	IDF1↑	MOTA↑	AssA↑	DetA↑
1.00	1.00	54.4	48.7	54.6	46.9	63.1
0.70	0.30	59.0	59.1	58.9	52.8	65.9
0.30	0.15	61.4	68.4	60.7	57.0	66.1
0.15	**0.06**	**62.8**	**71.1**	**62.3**	**57.9**	**68.8**
0.06	0.15	62.4	70.7	62.0	57.5	67.7
0.05	0.02	62.2	70.8	62.1	57.4	67.4
0.00	0.00	61.3	67.8	62.1	55.9	67.2

Table 4: Comparison with state-of-the-art methods on the MOT16 test set.

Method	HOTA↑	IDF1↑	MOTA↑	AssA↑	DetA↑
UTM You et al. (2023)	53.1	67.1	63.8	54.0	52.4
MeMOT Cai et al. (2022)	57.4	69.7	72.6	55.7	–
UCLS Meng et al. (2023)	58.4	70.4	73.0	–	–
OUTrack Liu et al. (2022)	59.2	71.1	74.2	57.0	61.8
GLMBFairMOT Van Ma et al. (2024)	59.4	72.4	75.0	57.7	61.5
Decode-MOT Lee et al. (2023)	60.2	73.0	74.7	59.0	61.7
SGT Hyun et al. (2023)	61.2	73.5	76.8	59.4	63.4
SCGTracker Zhang et al. (2024)	–	73.1	74.6	–	–
SR-MOT Li et al. (2024)	61.8	75.7	74.0	62.4	61.5
PAMBTrack Lv et al. (2026)	63.2	77.8	77.9	63.0	63.7
MTrack Yu et al. (2022)	–	77.8	78.5	–	–
FastTracker Hashempoor & Hwang (2025)	66.0	81.0	79.1	66.6	65.7
SpecHyper-Track	**66.5**	**81.9**	**82.5**	**66.9**	**66.8**

Effect of dual-frequency association. Next, we evaluate the frequency decomposition and stage-aware feature-routing strategy. Table 2 compares the base tracker, relation enhancement, two intermediate framelet implementations, and the dual-frequency association design. Dual-relation enhancement improves association quality over the base tracker, whereas the intermediate framelet

implementations yield lower HOTA and IDF1 than relation enhancement alone. Stage-aware routing improves association performance, with high-frequency-enhanced features used in stage 1 and fused features in stage 2. Compared with (f), the full model in (g) incorporates GPU-oriented implementation adjustments and conditionally applies hypergraph enhancement to reduce computation. The full model achieves 69.3 HOTA, 85.7 IDF1, and 73.2 AssA.

Sensitivity to relation weights. We analyze the homogeneous and heterogeneous residual weights α and β. As shown in Table 3, overly large α and β degrade performance by excessively perturbing identity-specific appearance features. The best balance is obtained at $\alpha = 0.15$ and $\beta = 0.06$, while removing both relation corrections consistently reduces association accuracy. These results support the use of moderate, asymmetric residual enhancement.

3.2 Comparison with State-of-the-art Methods

Tables 4-6 compare the proposed SpecHyper-Track with previous trackers on MOT16, MOT17, and DanceTrack under the corresponding benchmark protocols:

MOT16 and MOT17 Test Sets. As shown in Tables 4 and 5, our method achieves strong association performance on both pedestrian benchmarks, obtaining 66.5/66.9 HOTA and 66.9/67.7 AssA on MOT16/MOT17, respectively. On MOT17, it surpasses the strongest competing result in HOTA and AssA while maintaining competitive IDF1 and MOTA. These results indicate that relation-specific spectral reasoning improves identity association without sacrificing overall tracking accuracy.

DanceTrack Test Set. As shown in Table 6, on the more association-challenging DanceTrack benchmark, our method achieves 64.2 HOTA, 66.0 IDF1, and 49.1 AssA, outperforming the strongest competing online results in the table. The larger gains in association-oriented metrics indicate that the proposed relation-specific and stage-aware spectral modeling is particularly effective under similar appearances and complex motion.

Table 5: Comparison with state-of-the-art methods on the MOT17 test set.

Method	HOTA↑	IDF1↑	MOTA↑	AssA↑	DetA↑
DeepSORT+AAM Li et al. (2025b)	50.1	60.9	61.6	47.3	–
TrackFormer++ Liu et al. (2025a)	–	60.7	63.9	–	–
TrackSSM Hu et al. (2026)	61.4	74.1	78.5	59.6	63.6
OmniTracker Wang et al. (2025)	62.3	75.6	79.1	–	–
GHOST Seidenschwarz et al. (2023)	62.8	77.1	78.9	–	–
C-TWiX Miah et al. (2025)	63.1	76.3	78.1	62.5	64.1
TOPIC Cao et al. (2025)	63.9	78.6	78.8	64.3	63.7
PMTrack Guo et al. (2024)	64.0	80.2	78.9	63.6	–
C-BIOU Yang et al. (2023)	64.1	79.7	81.1	63.7	64.8
FineTrack Ren et al. (2023)	64.3	79.5	80.0	64.5	–
StrongSORT++ Du et al. (2023)	64.4	79.5	79.6	64.4	64.6
Deep OC-SORT Maggiolino et al. (2023)	64.9	80.6	79.4	65.9	–
DeconfuseTrack Huang et al. (2024a)	64.9	80.6	80.4	65.1	65.0
SparseTrack Liu et al. (2025b)	65.1	80.1	81.0	65.1	65.3
DetTrack Zhang et al. (2026)	65.2	–	80.6	65.7	65.1
LG-Track Meng et al. (2025)	65.4	80.4	81.4	65.4	65.6
CMTrack Shim et al. (2024a)	65.5	81.5	80.7	66.1	–
AdapTrack Shim et al. (2024b)	65.7	82.3	79.9	66.9	–
UCMCTrack+ Yi et al. (2024)	65.8	81.0	80.6	66.4	65.3
TrackTrack† Shim et al. (2025)	65.6	81.6	80.1	66.5	64.9
SpecHyper-Track	**66.9**	**82.4**	**81.7**	**67.7**	**66.4**

Table 6: Comparison with state-of-the-art methods on the DanceTrack test set.

Method	HOTA↑	IDF1↑	MOTA↑	AssA↑	DetA↑
ByteTrack Zhang et al. (2022)	47.1	51.9	88.2	31.5	70.5
DiffusionTrack Luo et al. (2024)	52.4	47.5	90.5	33.5	82.3
OC-SORT Cao et al. (2023)	55.1	54.6	92.0	38.3	–
SparseTrack Liu et al. (2025b)	55.5	58.3	91.3	39.1	–
ScoreMOT Zhao et al. (2024)	55.5	55.2	92.7	38.2	81.0
StrongSORT++ Du et al. (2023)	55.6	55.2	91.1	38.6	80.7
UTrack Solano-Carrillo et al. (2024)	55.8	56.4	89.7	39.3	79.4
DualTTrack Li et al. (2026a)	55.8	57.6	91.6	39.2	79.6
SH-Tracker Jia et al. (2026)	58.8	57.9	92.2	42.5	81.8
Preformer MOT Wang et al. (2026)	59.0	57.8	91.5	42.6	82.1
GeneralTrack Qin et al. (2024)	59.2	59.7	91.8	42.8	82.0
CBIoU Yang et al. (2023)	60.6	61.6	91.6	45.4	81.3
Deep OC-SORT Maggiolino et al. (2023)	61.3	61.5	92.3	45.8	82.2
CMTrack Shim et al. (2024a)	61.8	63.3	92.5	46.4	–
Hybrid-SORT Yang et al. (2024)	62.2	63.0	91.6	–	–
DiffMOT Lv et al. (2024)	62.3	63.0	92.8	47.2	82.5
TrackTrack† Shim et al. (2025)	62.4	64.7	93.6	47.0	83.2
SpecHyper-Track	**64.2**	**66.0**	**93.7**	**49.1**	**83.8**

† Reproduced by us using the same detector, ReID model, tracking configuration, hardware, and evaluation protocol as SpecHyper-Track.

4 Conclusion and Future Work

This work introduces a relation-specific spectral hypergraph framework for online multi-object tracking. Homogeneous and heterogeneous hyperedges model cooperative and competitive interactions, while low- and high-frequency operators capture shared context and identity-specific cues. Tight-framelet reconstruction provides aligned features for stage-aware association, with base features retained for trajectory memory. Experiments on MOT16, MOT17, and DanceTrack show improved association in crowded and appearance-ambiguous scenes.

The method remains sensitive to inaccurate motion estimates and noisy observations, and hypergraph-framelet operations add computational cost in dense scenes. Future work will focus on dynamic relation construction, adaptive frequency routing, and more efficient spectral operators.

AI DISCLOSURE

We used generative AI tools to assist with language editing and to support literature search and discovery of related work. All retrieved references were manually checked, and all technical content, experimental design, results, and claims were developed and verified by the authors. We take full responsibility for the final content of this work.

REFERENCES

Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 3464–3468. IEEE, 2016. doi: 10.1109/ICIP.2016.7533003.

Guillem Brasó and Laura Leal-Taixé. Learning a neural solver for multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6247–6257, 2020.

Jiarui Cai, Mingze Xu, Wei Li, Yuanjun 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*, 2022.

Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, 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*, pp. 9686–9696, 2023. doi: 10.1109/CVPR52729.2023.00934.

Wei Cao, Rachid Hedjam, Bessam Abdulrazak, and Jun-Peng Zhu. Poly-wavegc: A generalized spectral wavelet graph convolution network with adaptive orthogonal polynomials. In *Proceedings of the 39th Canadian Conference on Artificial Intelligence*, volume 318 of *Proceedings of Machine Learning Research*, pp. 342–353, 2026.

Xiaoyan Cao, Yiyao Zheng, Yao Yao, Huapeng Qin, Xiaoyu Cao, and Shihui Guo. Topic: A parallel association paradigm for multi-object tracking under complex motions and diverse scenes. *IEEE Transactions on Image Processing*, 34:743–758, 2025. doi: 10.1109/TIP.2025.3526066.

Zhoujuan Cui, Yuqi Dai, Yiping Duan, and Xiaoming Tao. Joint object detection and multi-object tracking based on hypergraph matching. *Applied Sciences*, 14(23):11098, 2024. doi: 10.3390/app142311098.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. Strongsort: Make deepsort great again. *IEEE Transactions on Multimedia*, 25:8725–8737, 2023.

Hantong Feng, Yonggang Wu, Duxin Chen, and Wenwu Yu. Tfwaveformer: Temporal-frequency collaborative multi-level wavelet transformer for dynamic link prediction. In *Proceedings of the ACM Web Conference*, pp. 4668–4679, 2026. doi: 10.1145/3774904.3792422.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3558–3565, 2019. doi: 10.1609/aaai.v33i01.33013558.

Shijie Guo et al. Pmtrack: Multi-object tracking with motion-aware. In *Proceedings of the Asian Conference on Computer Vision*, 2024.

Hamidreza Hashempoor and Yu Dong Hwang. Fasttracker: Real-time and accurate visual tracking. *arXiv preprint arXiv:2508.14370*, 2025.

Jiawei He, Zehao Huang, Naiyan Wang, and Zhaoxiang Zhang. Learnable graph matching: Incorporating graph partitioning with deep feature learning for multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5299–5309, 2021.

Bin Hu, Run Luo, Zelin Liu, Cheng Wang, and Wenyu Liu. Trackssm: A general motion predictor using a state space model. *Science China Information Sciences*, 69(5):152105, 2026. doi: 10.1007/s11432-024-4849-2.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. Heterogeneous graph transformer. In *Proceedings of The Web Conference*, pp. 2704–2710, 2020. doi: 10.1145/3366423.3380027.

Cheng Huang, Shoudong Han, Mengyu He, Wenbo Zheng, and Yuhao Wei. Deconfusetrack: Dealing with confusion for multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19290–19299, June 2024a.

Cheng Huang, Shoudong Han, Mengyu He, Wenbo Zheng, and Yuhao 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, 2024b.

Jing Huang and Jie Yang. UniGNN: A unified framework for graph and hypergraph neural networks. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 2563–2569, 2021. doi: 10.24963/ijcai.2021/353.

Jeongseok Hyun, Myunggu Kang, Dongyoon Wee, and Dit-Yan Yeung. Detection recovery in online multi-object tracking with sparse graph tracker. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4850–4859, 2023.

Yunhua Jia, Qinying Zhou, Chengjiang Zhou, and Yang Yang. Sh-tracker: Spatial hierarchy and virtual depth fusion for occlusion-aware multi-object tracking. *Results in Engineering*, 30:109675, 2026. doi: 10.1016/j.rineng.2026.109675.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. doi: 10.48550/arXiv.1609.02907.

Seong-Ho Lee, Dae-Hyeon Park, and Seung-Hwan Bae. Decode-mot: How can we hurdle frames to go beyond tracking-by-detection? *IEEE Transactions on Image Processing*, 32:4378–4392, 2023. doi: 10.1109/TIP.2023.3298538.

He Li, Zichen Wang, Weihang Kong, and Xingchen Zhang. Dualttrack: Improving information flow consistency in multi-object tracking via dual-track matching and recovery. *Information Processing & Management*, 63(8):104843, 2026a. doi: 10.1016/j.ipm.2026.104843.

Ming Li, Yujie Fang, Yi Wang, Han Feng, Yongchun Gu, Lu Bai, and Pietro Liò. Deep hypergraph neural networks with tight framelets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18385–18392, 2025a. doi: 10.1609/aaai.v39i17.34023.

Ming Li, Yongqi Li, Yuting Chen, Feilong Cao, and Ke Lv. Heterophily-aware contrastive learning for heterophilic hypergraphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23071–23078, 2026b. doi: 10.1609/aaai.v40i27.39473.

Xiang Li et al. Self-supervised re-identification for online joint multi-object tracking. *Knowledge and Information Systems*, 2024. doi: 10.1007/s10115-024-02237-w.

Yunhao Li, Zhen Xiao, Lin Yang, Dan Meng, Xin Zhou, Heng Fan, and Libo Zhang. Attmot: Improving multiple-object tracking by introducing auxiliary pedestrian attributes. *IEEE Transactions on Neural Networks and Learning Systems*, 36(3):5454–5468, 2025b. doi: 10.1109/TNNLS.2024.3384446.

Qiankun Liu, Dongdong Chen, Qi Chu, Lu Yuan, Bin Liu, Lei Zhang, and Nenghai Yu. Online multi-object tracking with unsupervised re-identification learning and occlusion estimation. *Neurocomputing*, 483:333–347, 2022. doi: 10.1016/j.neucom.2022.01.008.

Xinlong Liu, Kai Pu, Yunfeng Ping, Xiangli Yang, Zhisheng Yin, Zhangli Lan, and Bin Chen. Trackformer with prior position embedding and reference point updating for multiple object tracking. *IEEE Internet of Things Journal*, 12(18):37962–37973, 2025a. doi: 10.1109/JIOT.2025.3585134.

Zelin Liu, Xinggang Wang, Cheng 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*, 2025b.

Sitao Luan, Chenqing Hua, Qincheng Lu, Jiaqi Zhu, Mingde Zhao, Shuyuan Zhang, Xiao-Wen Chang, and Doina Precup. Revisiting heterophily for graph neural networks. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

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*, 129(2):548–578, 2021.

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. doi: 10.1609/aaai.v38i5.28192.

Wenhan Luo, Junliang Xing, Anton Milan, Xiaoqin Zhang, Wei Liu, and Tae-Kyun Kim. Multiple object tracking: A literature review. *Artificial Intelligence*, 293:103448, 2021. doi: 10.1016/j.artint.2020.103448.

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. doi: 10.1109/CVPR52733.2024.01828.

Zheng Lv, Tao Lu, Yanduo Zhang, and Jiaming Wang. Multi-object tracking with prediction association of motion behavior. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E109-A(4):806–817, 2026. doi: 10.1587/transfun.2025EAP1089.

Gerard Maggiolino, Adnan Ahmad, Jinkun Cao, and Kris Kitani. Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification. In *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 3025–3029. IEEE, 2023.

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.

Shuang Meng, Dong Shao, Jian Guo, and Shenghua Gao. Tracking without label: Unsupervised multiple object tracking via contrastive similarity learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16264–16273, 2023.

Ting Meng, Chunyun Fu, Mingguang Huang, Tao Huang, Xiyang Wang, Jiawei He, and Wankai Shi. Localization-guided track: A deep association multiobject tracking framework based on localization confidence of camera detections. *IEEE Sensors Journal*, 25(3):5282–5293, 2025. doi: 10.1109/JSEN.2024.3522021.

Mehdi Miah, Guillaume-Alexandre Bilodeau, and Nicolas Saunier. Learning data association for multi-object tracking using only coordinates. *Pattern Recognition*, 160:111169, 2025. doi: 10.1016/j.patcog.2024.111169.

Anton Milan, Konrad Schindler, and Stefan Roth. Detection- and trajectory-level exclusion in multiple object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3682–3689, 2013. doi: 10.1109/CVPR.2013.472.

Ziqi Pang, Jie Li, Pavel Tokmakov, Dian Chen, Sergey Zagoruyko, and Yu-Xiong Wang. Standing between past and future: Spatio-temporal modeling for multi-camera 3d multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17928–17938, 2023. doi: 10.1109/CVPR52729.2023.01719.

Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 261–268, 2009. doi: 10.1109/ICCV.2009.5459260.

Zheng Qin, Sanping Zhou, Le Wang, Jinghai Duan, Gang Hua, and Wei Tang. Towards generalizable multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Hao Ren et al. 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*, 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, Jubi Hwang, Kangwook Ko, and Changick Kim. A confidence-aware matching strategy for generalized multi-object tracking. In *2024 IEEE International Conference on Image Processing (ICIP)*, pp. 4042–4048, 2024a. doi: 10.1109/ICIP51287.2024.10647729.

Kyujin Shim, Kangwook Ko, Jubi Hwang, and Changick Kim. Adaptrack: Adaptive thresholding-based matching for multi-object tracking. In *2024 IEEE International Conference on Image Processing (ICIP)*, pp. 2222–2228, 2024b. doi: 10.1109/ICIP51287.2024.10648139.

Kyujin Shim, Kangwook Ko, Yujin 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. doi: 10.1109/CVPR52734.2025.01091.

Edgardo Solano-Carrillo, Felix Sattler, Antje Alex, Alexander Klein, Bruno Pereira Costa, Ángel Bueno Rodríguez, and Jannis Stoppe. Utrack: Multi-object tracking with uncertain detections. In *European Conference on Computer Vision Workshops*, pp. 219–236, 2024. doi: 10.1007/978-3-031-91585-7_14.

Zikai Song, Junqing Yu, Yi-Ping Phoebe Chen, Wei Yang, and Xinchao Wang. Hypergraph-state collaborative reasoning for multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28123–28133, 2026.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20993–21002, 2022.

Nguyen Van Ma et al. Glmbfairmot: An efficient multi-object tracking framework combining fairmot and the generalized labeled multi-bernoulli filter. *Pattern Recognition*, 2024.

Junke Wang, Zuxuan Wu, Dongdong Chen, Chong Luo, Xiyang Dai, Lu Yuan, and Yu-Gang Jiang. Omnitracker: Unifying visual object tracking by tracking-with-detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):3159–3174, 2025. doi: 10.1109/TPAMI.2025.3529926.

Yueying Wang, Yuhao Qing, Kaer Huang, Chuangyin Dang, and Zhengtian Wu. Preformer mot: A transformer-based approach for multi-object tracking with global trajectory prediction. *Fundamental Research*, 6(1):423–431, 2026. doi: 10.1016/j.fmre.2024.06.015.

Longyin Wen, Zhen Lei, Siwei Lyu, Stan Z. Li, and Ming-Hsuan Yang. Exploiting hierarchical dense structures on hypergraphs for multi-object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(10):1983–1996, 2016. doi: 10.1109/TPAMI.2015.2509979.

Longyin Wen, Dawei Du, Shengkun Li, Xiao Bian, and Siwei Lyu. Learning non-uniform hypergraph for multi-object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 8981–8988, 2019. doi: 10.1609/aaai.v33i01.33018981.

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*, pp. 3645–3649, 2017. doi: 10.1109/ICIP.2017.8296962.

Bingbing Xu, Huawei Shen, Qi Cao, Yunqi Qiu, and Xueqi Cheng. Graph wavelet neural network. In *International Conference on Learning Representations*, 2019.

Fan Yang, Shigeyuki Odashima, Shoichi Masui, and Shan Jiang. Hard to track objects with irregular motions and similar appearances? make it easier by buffering the matching space. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4799–4808, 2023.

Mingzhan Yang, Guangxin 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. doi: 10.1609/aaai.v38i7.28471.

Kefu Yi, Kai Luo, Xiaolei Luo, Jiangui Huang, Hao Wu, Rongdong Hu, and Wei Hao. Ucmctrack: Multi-object tracking with uniform camera motion compensation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Sisi You, Hantao Yao, Bing-Kun Bao, and Changsheng Xu. Utm: A unified multiple object tracking model with identity-aware feature enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21876–21886, June 2023.

En Yu, Zhuoling Li, and Shoudong Han. Towards discriminative representation: Multi-view trajectory contrastive learning for online multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8834–8843, June 2022.

Le Yu, Leilei Sun, Bowen Du, Chuanren Liu, Weifeng Lv, and Hui Xiong. Heterogeneous graph representation learning with relation awareness. *IEEE Transactions on Knowledge and Data Engineering*, 35(6):5935–5947, 2023. doi: 10.1109/TKDE.2022.3160208.

Shuaicheng Zhang, Haohui Wang, Junhong Lin, Xiaojie Guo, Yada Zhu, Si Zhang, Dongqi Fu, and Dawei Zhou. HeroFilter: Adaptive spectral graph filter for varying heterophilic relations. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Yajuan Zhang, Yongquan Liang, Jiaxu Leng, and Zhihui Wang. Scgtracker: Spatio-temporal correlation and graph neural networks for multiple object tracking. *Pattern Recognition*, 149:110249, 2024. doi: 10.1016/j.patcog.2023.110249.

Yi Zhang, Yi Su, and Chen Luo. Dettrack: Realizing strong identity preservation in multi-object tracking via exploration of detection information. In *ECAI 2025*, 2026. doi: 10.3233/FAIA250837.

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 *European Conference on Computer Vision*, pp. 1–21, 2022. doi: 10.1007/978-3-031-20047-2_1.

Zhen Zhao et al. ScoreMOT: A score-based multi-object tracking method. *Image and Vision Computing*, 2024.

Shuai Zheng, Zhenfeng Zhu, Zhizhe Liu, Youru Li, and Yao Zhao. Node-oriented spectral filtering for graph neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46 (1):388–402, 2024. doi: 10.1109/TPAMI.2023.3324937.

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.

Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

A APPENDIX

This appendix provides extended discussions of related work, theoretical analysis, implementation details, and supplementary experiments for SpecHyper-Track. We first review prior work on multi-object tracking, graph- and hypergraph-based relation modeling, and frequency-aware graph and hypergraph learning, clarifying their connections to and differences from our approach. We then analyze the spectral properties of the relation-specific operators, the frequency-selective interpretation of framelet residuals, and the boundedness of the residual correction before normalization. The remaining sections present computational complexity, experimental configurations and limitations, followed by qualitative tracking results, cross-dataset performance comparisons, and affinity heatmaps.

A.1 RELATED WORK

A.1.1 DETECTION-BASED ONLINE MULTI-OBJECT TRACKING

Online tracking-by-detection (TBD) methods associate frame-level detections with existing trajectories using motion, appearance, and confidence cues. SORT combines Kalman filtering with bipartite matching to provide an efficient motion-based baseline Bewley et al. (2016). DeepSORT introduces discriminative appearance embeddings to improve identity preservation under occlusion Wojke et al. (2017). ByteTrack further recovers potential targets by associating low-confidence detections Zhang et al. (2022). Beyond conventional TBD pipelines, TrackFormer jointly performs detection and association through identity-preserving track queries Meinhardt et al. (2022).

Despite these advances, most online MOT frameworks still formulate association primarily as trajectory–detection matching, where the compatibility between individual trajectories and detections is evaluated independently. Such pairwise association paradigms provide limited capability to capture structured interactions among multiple targets in crowded scenes. For example, objects may exhibit cooperative motion patterns that provide shared contextual evidence, while visually similar targets may introduce competitive interactions requiring stronger identity discrimination. These limitations motivate the exploration of high-order relational modeling beyond conventional pairwise association.

A.1.2 GRAPH- AND HYPERGRAPH-BASED RELATION MODELING

Graph-based trackers extend independent trajectory–detection matching by explicitly modeling contextual dependencies among tracking entities. MPNTrack performs global data association through spatio-temporal message passing over a detection graph Brasó & Leal-Taixé (2020). GMTracker formulates association as learnable graph matching and exploits second-order structural consistency between tracklet and detection graphs He et al. (2021). Hypergraphs further generalize pairwise edges to group-wise connections, enabling high-order interactions among multiple tracking entities. FH^{2}T represents tracklet affinities using hierarchical dense hypergraph structures to improve tracking under occlusion and appearance ambiguity Wen et al. (2016). The non-uniform hypergraph tracker learns hyperedges of different cardinalities to capture dependencies of varying orders Wen et al. (2019). JDTHM integrates hypergraph matching with object detection and historical feature aggregation in an end-to-end framework Cui et al. (2024). More recently, HyperSSM combines dynamic hypergraph reasoning with state-space modeling for collaborative motion estimation Song et al. (2026). Collectively, these studies demonstrate that high-order connectivity can provide contextual constraints beyond conventional pairwise association.

However, existing hypergraph-based MOT methods mainly emphasize richer connectivity or stronger representation learning, while the semantic differences among hyperedges are rarely reflected in the propagation process. In crowded scenes, motion-consistent targets form cooperative relations that benefit from contextual aggregation, whereas nearby targets from different motion groups form competitive relations that require stronger identity separation. This motivates us to distinguish interaction semantics explicitly rather than treating all hyperedges uniformly. Accordingly, we introduce edge-semantic heterogeneous hypergraph modeling, where homogeneous hyperedges represent cooperative intra-group relations and heterogeneous hyperedges capture competitive inter-group interactions. These two relation types are further associated with relation-specific spectral operators and stage-aware association.

A.1.3 Frequency-Aware Graph and Hypergraph Learning

Spectral graph learning decomposes graph signals with respect to the graph Laplacian, enabling frequency-selective representation learning. ChebNet approximates localized spectral filters using Chebyshev polynomials, avoiding explicit eigendecomposition Defferrard et al. (2016). GWNN replaces Fourier bases with localized graph wavelets to capture multi-scale structural patterns efficiently Xu et al. (2019). Poly-WaveGC further constructs adaptive wavelet filters using learnable orthogonal polynomial bases Cao et al. (2026). For high-order relational data, FrameHGNN introduces tight-framelet hypergraph convolutions that jointly model low- and high-frequency components Li et al. (2025a). TFWaveFormer integrates multi-resolution wavelet decomposition with temporal attention for dynamic graph representation learning Feng et al. (2026). These studies demonstrate that frequency-aware decomposition provides complementary structural and temporal information beyond conventional spatial aggregation.

However, existing approaches typically apply shared spectral transformations without explicitly coupling frequency processing with relation semantics or trajectory states. This is restrictive for online MOT, where cooperative and competitive interactions exhibit different spectral characteristics, and trajectories at different lifecycle stages require different representations. Established trajectories benefit more from discriminative high-frequency cues, whereas newly initialized trajectories require stable contextual information together with identity-sensitive features. Accordingly, we introduce relation-specific spectral operators on semantically distinct hypergraphs and adaptively route low- and high-frequency representations according to trajectory states, integrating spectral reasoning directly into the cascaded association process.

A.2 Theoretical Analysis of Relation-Specific Spectral Propagation

This section provides theoretical support for the proposed relation-specific spectral refinement. We first establish the spectral bounds of the normalized hypergraph operator, which underpin the frequency interpretation of our propagation rules. We then characterize the complementary smoothing and sharpening effects of the two relation branches and explain the frequency-selective role of the framelet residuals under the tight-frame assumption. Finally, we show that the relation-specific residual correction is bounded before normalization, clarifying how the update constrains perturbations to the original appearance features.

Lemma 1 (Spectrum of the normalized hypergraph operator). For a relation-specific hypergraph with nonnegative hyperedge weights, the normalized propagation operator

$$\Theta_r = (D_v^r)^{-1/2} H^r W_e^r (D_e^r)^{-1} (H^r)^\top (D_v^r)^{-1/2}$$

satisfies

$$0 \preceq \Theta_r \preceq I.$$

Consequently, all eigenvalues of $L_r = I - \Theta_r$ lie in $[0, 1]$.

Proof. Let

$$B_r = (D_v^r)^{-1/2} H^r (W_e^r)^{1/2} (D_e^r)^{-1/2}.$$

Then $\Theta_r = B_r B_r^\top$, and hence $\Theta_r \succeq 0$. Let $d_i^r = (D_v^r)_{ii}$ denote the degree of node i, and let $\delta_e^r = |e|$ and w_e^r denote the cardinality and weight of hyperedge e, respectively. Following the convention in the main text, $(d_i^r)^{-1/2}$ is set to zero whenever $d_i^r = 0$. For any vector x,

$$x^\top \Theta_r x = \sum_e \frac{w_e^r}{\delta_e^r} \left(\sum_{i \in e} \frac{x_i}{\sqrt{d_i^r}} \right)^2.$$

By the Cauchy–Schwarz inequality,

$$\left(\sum_{i \in e} \frac{x_i}{\sqrt{d_i^r}} \right)^2 \leq \delta_e^r \sum_{i \in e} \frac{x_i^2}{d_i^r}.$$

Therefore,

$$x^\top \Theta_r x \leq \sum_{i : d_i^r > 0} \frac{x_i^2}{d_i^r} \sum_{e \ni i} w_e^r = \sum_{i : d_i^r > 0} x_i^2 \leq \|x\|_2^2.$$

Thus, $\Theta_r \preceq I$, and the spectrum of $L_r = I - \Theta_r$ is contained in $[0, 1]$. □

Proposition 1 (Relation-specific spectral response). Let $L_r = I - \Theta_r$ be the normalized hypergraph Laplacian with eigenpair $(\lambda_{r,i}, u_{r,i})$. Ignoring the nonlinear bounding, relation masks, and row-wise normalization, and considering each relation branch separately, the homogeneous and heterogeneous updates in Eq. (13) have spectral responses

$$h_h(\lambda) = 1 - \alpha\lambda, \qquad h_g(\lambda) = 1 + \beta\lambda.$$

For $0 < \alpha \leq 1$ and $\beta > 0$, $h_h(\lambda)$ decreases with λ, whereas $h_g(\lambda)$ increases with λ. Therefore, the homogeneous branch suppresses high-frequency variations, while the heterogeneous branch emphasizes them.

Proof. For an eigenvector $u_{r,i}$ satisfying $L_r u_{r,i} = \lambda_{r,i} u_{r,i}$, we have

$$(I - \alpha L_h)u_{h,i} = (1 - \alpha\lambda_{h,i})u_{h,i},$$

and

$$(I + \beta L_g)u_{g,i} = (1 + \beta\lambda_{g,i})u_{g,i}.$$

Thus, the homogeneous response decreases with frequency, whereas the heterogeneous response increases with frequency. □

Proposition 2 (Frequency-selective residual interpretation). For each relation-specific hypergraph $r \in \{h, g\}$, assume that the corresponding framelet operators form a tight frame, such that

$$W_{(0,J,r)}^\top W_{(0,J,r)} + \sum_{j=1}^{J} W_{(1,j,r)}^\top W_{(1,j,r)} = I.$$

For the homogeneous relation, the low-frequency reconstruction $F_l = W_{(0,J,h)}^\top W_{(0,J,h)} X_b$ satisfies

$$F_l - X_b = -\sum_{j=1}^{J} W_{(1,j,h)}^\top W_{(1,j,h)} X_b.$$

Hence, the residual term $F_l - X_b$ suppresses the high-frequency components associated with homogeneous relations. In contrast, the heterogeneous reconstruction

$$F_h = \sum_{j=1}^{J} W_{(1,j,g)}^\top W_{(1,j,g)} X_b$$

directly aggregates high-frequency components and therefore reinforces variations across heterogeneous relations.

Proof. From the tight-frame reconstruction identity,

$$X_b = W_{(0,J,h)}^\top W_{(0,J,h)} X_b + \sum_{j=1}^{J} W_{(1,j,h)}^\top W_{(1,j,h)} X_b.$$

Substituting the definition of F_l and rearranging gives

$$F_l - X_b = -\sum_{j=1}^{J} W_{(1,j,h)}^\top W_{(1,j,h)} X_b.$$

The heterogeneous result follows directly from the definition of F_h. □

The identities above hold under the exact tight-frame assumption; with the finite-order Chebyshev approximation used in the implementation, they hold up to approximation error.

Proposition 3 (Bounded relation-specific perturbation). Let

$$\Delta_i = \alpha m_i^h \tanh(((\Theta_h - I)X)_i) + \beta m_i^g \tanh(((I - \Theta_g)X)_i)$$

denote the pre-normalization residual correction of node i in Eq. (13). For $\alpha, \beta \geq 0$ and a d-dimensional node representation,

$$\|\Delta_i\|_2 \leq \sqrt{d}\,\left(\alpha m_i^h + \beta m_i^g\right).$$

Therefore, the relation-specific residual correction introduces a bounded perturbation to the original appearance representation before normalization.

Proof. Since $\tanh(\cdot)$ is applied element-wise and satisfies $|\tanh(x)| \leq 1$, each d-dimensional residual vector has Euclidean norm at most $\sqrt{d}$. By the triangle inequality,

$$\|\Delta_i\|_2 \leq \alpha m_i^h \sqrt{d} + \beta m_i^g \sqrt{d},$$

which yields the stated bound. □

This bound controls the magnitude of the residual correction before normalization, which is particularly important when hyperedges contain noisy or ambiguous associations.

A.3 Complexity Analysis

Let N, M, K, and d denote the number of nodes, the total number of hyperedges, the maximum hyperedge order, and the feature dimension, respectively. Semantic relation discovery evaluates pairwise motion, appearance, and spatial compatibility with a worst-case complexity of $\mathcal{O}(N^2 d)$. In practice, spatial gating and the maximum hyperedge order reduce the number of candidate relations retained for subsequent high-order reasoning.

The proposed spectral formulation implements frequency filtering through polynomial functions of relation-specific Laplacians rather than repeated explicit spectral transforms. With sparse incidence matrices and implicit application of the propagation operators, the cost of relation-specific propagation scales approximately with the number of nonzero node–hyperedge incidences, yielding $\mathcal{O}(MKd)$. With a K_c-order Chebyshev approximation, the multi-frequency filtering stage requires approximately $\mathcal{O}(K_c MKd)$. Since K and K_c are small fixed constants, the dominant spectral cost grows approximately linearly with the number of retained hyperedges and feature dimension.

Compared with eigenvector-based spectral filtering, this formulation replaces dense global transforms with localized polynomial propagation, which is better suited to online MOT where the relational structure is rebuilt at each frame. For isolated nodes or scenes without valid relations, the corresponding relation mask becomes zero, so no additional spectral correction is introduced.

A.4 Additional Experimental Details

We follow the official data splits and evaluation protocols. For MOT16 and MOT17, we use YOLOX-generated detections under the private-detection protocol. For DanceTrack, results are reported on the official test set. The detector, ReID extractor, and basic tracking configurations are fixed across all experimental variants.

For MOT16 and MOT17, experiments are conducted under the private-detection protocol using YOLOX-x detections. The SBS50 model implemented in FastReID is used for appearance extraction, producing a 2048-dimensional ℓ_2-normalized descriptor for each detection and trajectory. The detector, ReID network, and basic tracking configurations remain fixed across all experimental variants, and the proposed spectral enhancement module requires no additional training. Following the association cost, the weights are set to $\omega_{\text{iou}} = 0.50$, $\omega_{\text{app}} = 0.50$, $\omega_{\text{conf}} = 0.10$, and $\omega_{\text{ang}} = 0.05$.

At frame t, the hypergraph nodes consist of current detections, previous-frame contextual nodes, and Kalman-predicted trajectories. For motion grouping, the direction-similarity, spatial-proximity, and normalized distance terms are weighted by $\omega_m = 0.45$, $\omega_s = 0.35$, and $\omega_d = 0.20$, respectively. The grouping threshold, maximum spatial radius, and maximum hyperedge order are set to $\theta_g = 0.55$, $r_{\max} = 120$ pixels, and $K_{\max} = 8$, respectively.

For cross-group heterogeneous relations, appearance similarity and spatial proximity are weighted by $\lambda_a = 0.65$ and $\lambda_s = 0.35$. Following the cross-group candidate criterion, the search radius r_h, appearance-ambiguity threshold θ_a, and close-distance threshold r_c are set to 180 pixels, 0.45, and 60 pixels, respectively.

The homogeneous and heterogeneous residual weights are set to $\alpha = 0.15$ and $\beta = 0.06$. We employ Haar tight framelets with two decomposition levels and a sixth-order Chebyshev approximation, i.e., $J = 2$ and $K_c = 6$. The low- and high-frequency residual coefficients are set to $\gamma_l = 0.15$ and $\gamma_h = 0.40$, while the fusion weight is fixed to $\eta = 0.5$. Stage 1 of association

uses the high-frequency representation X_h, whereas stage 2 uses the fused representation X_f. After association, the base representation X_b is restored for trajectory-memory updating.

In the full model, hypergraph enhancement is applied only when at least eight detections are present and at least one detection pair has a center distance below 120 pixels.

The method is implemented using PyTorch, NumPy, and SciPy and evaluated on a laptop equipped with an NVIDIA GeForce GTX 1650 GPU with 4 GB dedicated memory and 16 GB system RAM. Unless otherwise specified, all hyperparameters are fixed across MOT16, MOT17, and DanceTrack.

A.5 Limitations

Despite the consistent improvements across different benchmarks, the proposed method still has several limitations.

(1) The estimated motion relations may become unreliable in complex interaction scenarios. When visually similar targets move in the same direction near an intersection, they may be assigned to the same motion group, causing low-frequency propagation to mix features from different identities. Abrupt turns, rapid acceleration, and strong camera motion can also reduce the accuracy of short-term velocity estimates and Kalman predictions, resulting in incorrect heterogeneous connections. Although residual enhancement and complete-link constraints alleviate these errors, they cannot fully eliminate identity confusion caused by inaccurate grouping.

(2) High-frequency responses are sensitive to noisy observations. Under severe occlusion, low resolution, motion blur, or unstable detections, bounding-box jitter and unreliable ReID features may generate spurious high-frequency components, increasing the risk of mismatches and trajectory fragmentation. Long-term complete occlusion remains particularly challenging because predicted trajectories gradually drift without reliable contextual observations.

(3) The effectiveness and computational cost of high-order modeling vary with scene density. In sparse scenes, the limited number of neighboring targets provides insufficient high-order context, reducing the benefit of hypergraph enhancement. In extremely crowded scenes, the growing number and order of hyperedges increase the computational overhead of hypergraph construction and Framelet transforms, which may limit runtime scalability.

These limitations suggest directions for future extensions and may guide subsequent MOT designs in complex real-world environments.

A.6 Visualization and Heatmap Analysis

To provide further qualitative evidence, we present visualization results and affinity heatmap analyses on MOT17 and DanceTrack, together with a cross-dataset benchmark comparison. These examples complement the quantitative results in the main paper by illustrating how the proposed method maintains stable identities, resolves ambiguous associations, and improves relation modeling in crowded and dynamic scenes.

Figure 4 presents qualitative tracking results on representative sequences selected from the MOT16, MOT17, and DanceTrack datasets. These examples cover diverse tracking scenarios, including crowded pedestrian scenes, frequent occlusions, close interactions, and complex motion patterns. As shown in the figure, the proposed method can accurately localize targets with relatively precise bounding boxes while maintaining stable identity labels over time. Even when multiple targets are spatially close or partially occluded, most trajectories remain continuous and identity switches are effectively suppressed. The consistent tracking performance across the three datasets further demonstrates the robustness of the proposed method under different scene characteristics and motion distributions.

Figure 5 shows the benchmark comparison across MOT16, MOT17, and DanceTrack. By jointly visualizing AssA, IDF1, and HOTA, the figure indicates that the proposed method achieves a favorable trade-off between association accuracy and overall tracking quality across different benchmarks.

Figure 6 provides a qualitative case study on DanceTrack0003. The examples highlight three representative behaviors of the proposed method: stable identity preservation across consecutive frames,



Figure 4: Visualization of tracking results on the MOT16, MOT17, and DanceTrack datasets. Tracks are color-coded for clarity, and the proposed method maintains stable identities and continuous trajectories across diverse scenes with frequent occlusions and crowd interactions.

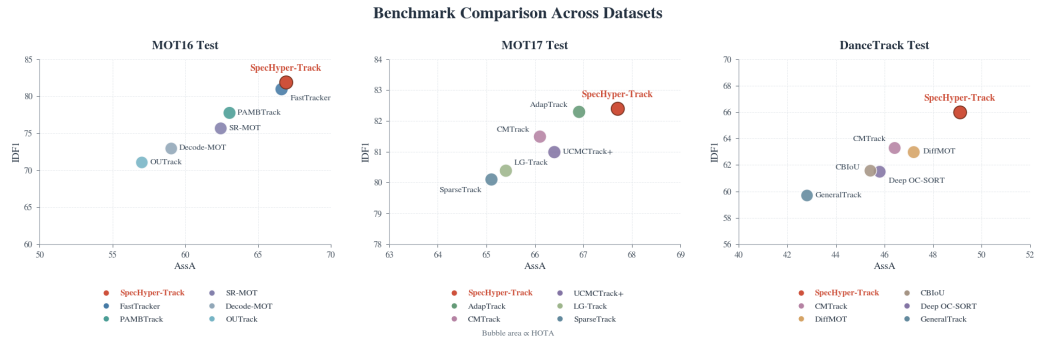


Figure 5: Performance comparison across MOT16, MOT17, and DanceTrack. The horizontal and vertical axes denote AssA and IDF1, respectively, while the bubble size is proportional to HOTA. SpecHyper-Track consistently achieves a favorable balance between association accuracy, identity preservation, and overall tracking performance across the three benchmarks.

robust association under crowded and highly overlapping interactions, and reliable tracking under substantial pose and motion variations.

Figure 7 presents additional qualitative results on MOT17-12. The selected frames demonstrate that the tracker remains effective under challenging factors such as scale variation, crowded interaction, cross-depth motion, and motion blur.

Figures 8 and 9 compare inter-target affinities before and after refinement on MOT17-12 and DanceTrack0003. The diagonal entries are masked. In the selected examples, refinement reduces off-diagonal affinities, indicating stronger separation between different targets.

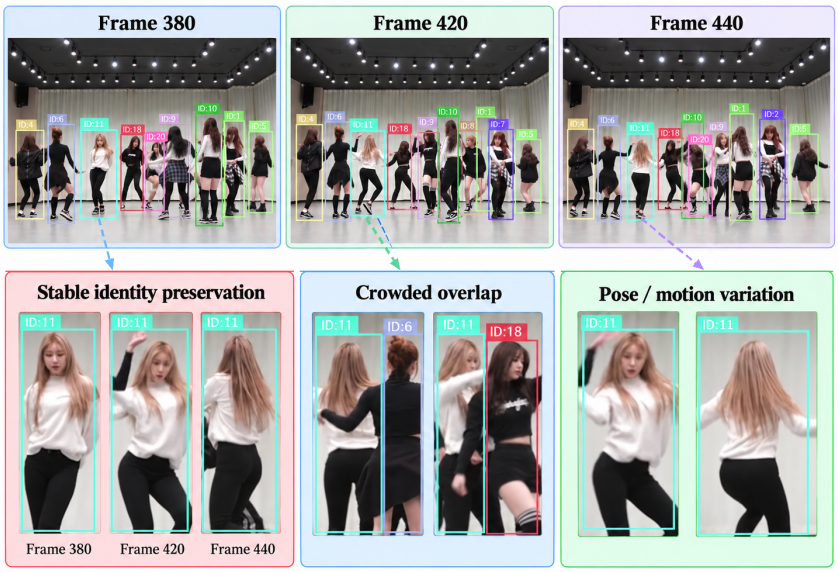


Figure 6: Qualitative results on DanceTrack0003, showing identity preservation across frames, crowded overlap, and pose changes.



Figure 7: Qualitative results on MOT17-12, showing the same target (ID 14) at different scales across frames 100, 150, and 200.

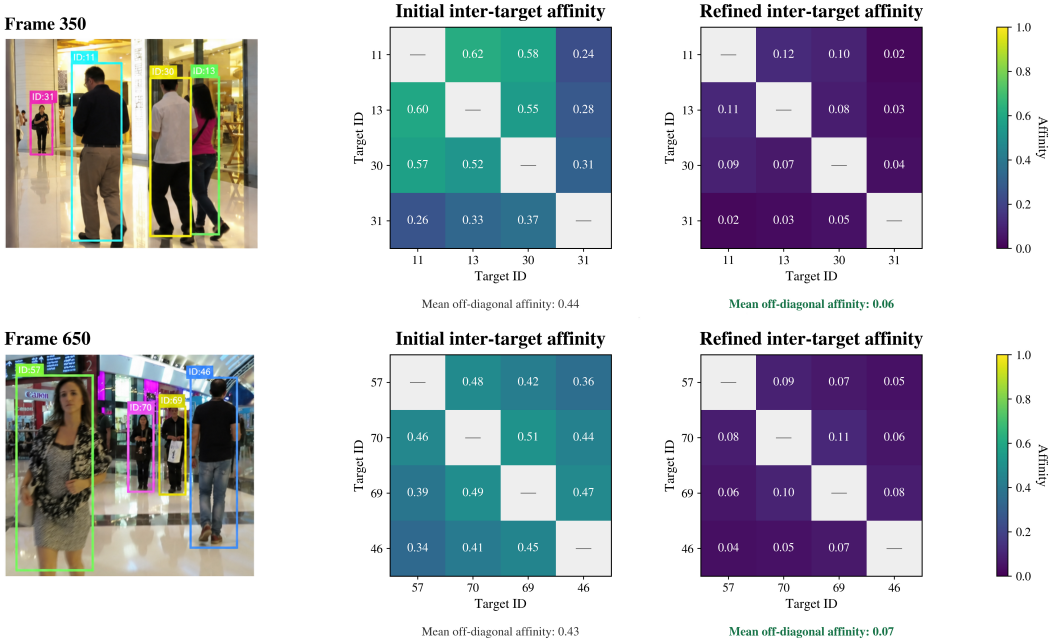


Figure 8: Inter-target affinity heatmaps on MOT17-12 before and after refinement. Off-diagonal affinities decrease in the selected examples, indicating stronger separation between different targets. Diagonal entries are masked.

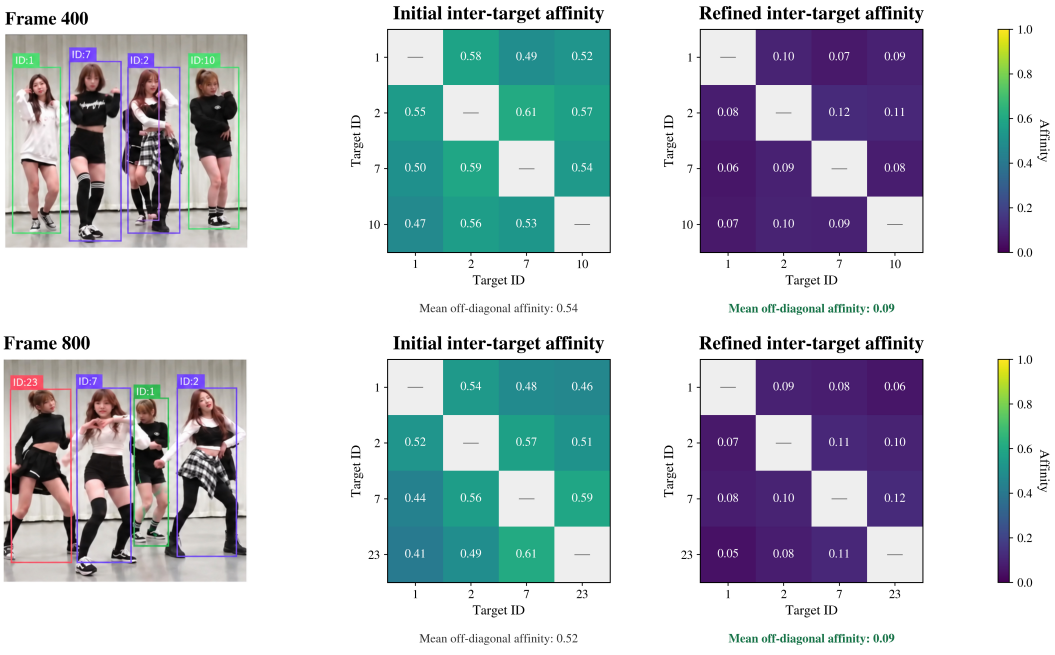


Figure 9: Inter-target affinity heatmaps on DanceTrack0003 before and after refinement. The selected examples show reduced affinities between different targets after refinement. Diagonal entries are masked.