COAL: Counterfactual and Observation-Enhanced Alignment Learning for Discriminative Referring Multi-Object Tracking

Shukun Jia[1,2], **Shiyu Hu**[3], **Yipei Wang**[1,2], **Ximeng Cheng**[1,2], **Yichao Cao**[4] and **Xiaobo Lu**[1,2,*]
[1]School of Automation, Southeast University, Nanjing, China
[2]Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Nanjing, China
[3]School of Physical & Mathematical Sciences, Nanyang Technological University, Singapore
[4]Big Data Institute, Central South University, Changsha, China
jia_shukun@seu.edu.cn, xblu2013@126.com

Abstract

Referring Multi-Object Tracking (RMOT) faces a fundamental structural contradiction between the high-discriminability demand and the sparse semantic supervision. This mismatch is particularly acute in highly homogeneous scenarios that require fine-grained discrimination over complex compositional semantics. However, under sparse supervision, models overfit to salient yet insufficient cues, leading to shortcut learning and degraded compositional discrimination. To resolve this, we propose **COAL** (**C**ounterfactual and **O**bservation-enhanced **A**lignment **L**earning), a framework that advances RMOT beyond isolated structural optimization through knowledge regularization. First, we introduce Explicit Semantic Injection (ESI) via a VLM to densify the observation space and enhance instance discriminability. Second, leveraging LLM reasoning, we propose Counterfactual Learning (CFL) to augment supervision, enforcing strict attribute verification for robust compositional recognition. These strategies are unified within a Hierarchical Multi-Stream Integration (HMSI) architecture, which distills external knowledge into domain-specific discriminative representations. Experiments on Refer-KITTI and Refer-KITTI-V2 benchmarks validate COAL's efficacy. Notably, it surpasses the state-of-the-art by 7.28% HOTA on the highly challenging Refer-KITTI-V2. These results demonstrate the effectiveness of knowledge regularization for resolving the sparsity–discriminability paradox in RMOT.

1 Introduction

Referring Multi-Object Tracking (RMOT) requires an agent to recognize and track arbitrary objects specified by natural language descriptions. Unlike conventional Multi-Object Tracking (MOT), which primarily operates within the visual space, RMOT fundamentally demands fine-grained cross-modal discrimination.

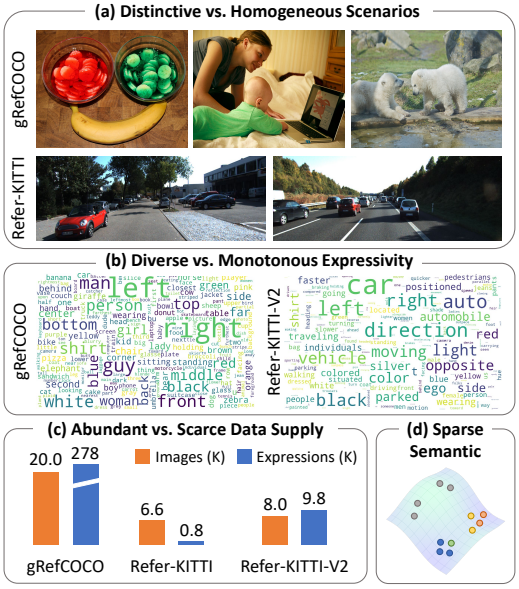

Figure 1: **The Sparsity-Discriminability Paradox in RMOT.** We illustrate the structural contradiction inherent to RMOT: The high-discriminability demand imposed by visual homogeneity (a) and restricted vocabulary (b) is confronted with the sparsely populated semantic manifold (d) resulting from the limited data scale (c).

Despite great advancements, existing methods [Zhuang *et al.*, 2025; Zhao *et al.*, 2025; Li *et al.*, 2025b] predominantly formulate RMOT as a multi-modal fusion problem, implicitly assuming that the discrimination bottleneck can be resolved in isolation by increasing interaction intricacy or architectural complexity. However, this perspective overlooks a more fundamental structural contradiction inherent to the task: the mismatch between the high-discriminability demand and the extremely sparse semantic supervision. Unlike general visual grounding [Liu *et al.*, 2023], this contradiction is particularly pronounced in RMOT, as illustrated in Fig. 1.

First, RMOT scenarios exhibit severe visual homogeneity (Fig. 1a). In representative benchmarks like Refer-KITTI

*Corresponding author. Supplementary material is available at: https://arxiv.org/abs/2605.14795.

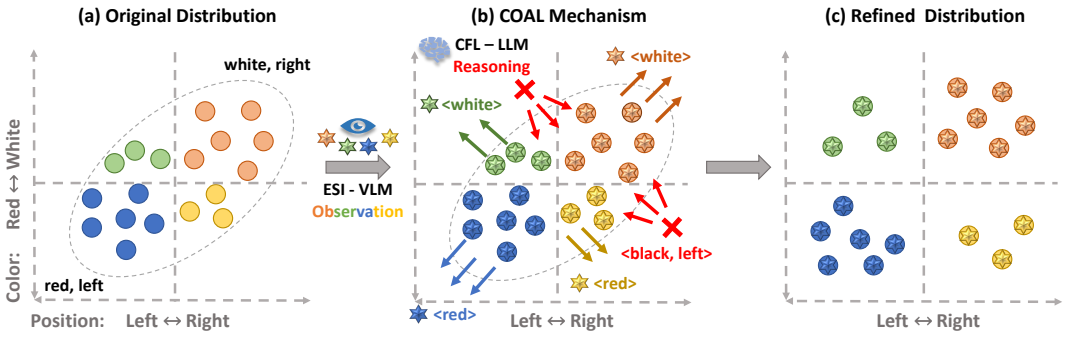

Figure 2: **Conceptual Mechanism of COAL.** Due to sparse semantic supervision, attributes tend to be coupled in the original distribution (a). ESI injects VLM-derived discriminative cues, denoted by star symbols, to enhance semantic separability, while CFL introduces LLM-derived counterfactual supervision, denoted by red crosses, to enforce compositional discrimination (b). Consequently, the refined representation becomes semantically discriminative and structurally comprehensive (c).

[Wu *et al.*, 2023] and Refer-KITTI-V2 [Zhang *et al.*, 2024], targets share high visual and semantic similarity with numerous homogeneous distractors. This renders coarse-grained discrimination strategies ineffective, necessitating the recognition of subtle, non-salient attributes. Second, RMOT expressions are characterized by restricted lexical composition (Fig. 1b), as reflected by the highly repetitive attribute combinations in benchmark annotations. Lacking unique keywords or rich contextual cues, the model is forced to distinguish targets by resolving the rigorous combination of limited attributes. For example, distinguishing "the red car on the left" from "the white car on the left" or "the red car on the right" requires accurate compositional interpretation beyond coarse attribute matching. Finally, these discrimination requirements face severe supervision scarcity: the limited data scale (Fig. 1c) leaves the vast combinatorial space of attributes severely undersampled, giving rise to a sparse semantic manifold (Fig. 1d).

The coexistence of fine-grained discrimination requirements and sparse supervision can induce a systematic perception bias. Models tend to prioritize easily learnable but insufficient attributes as primary discriminators, leading to attribute coupling and spurious correlations. Consequently, the model degenerates into shortcut learning, resulting in feature collapse. Fundamentally, this issue represents a breakdown in compositional generalization, rather than merely insufficient representation capacity or fusion architecture design.

To this end, we propose **COAL** (**C**ounterfactual and **O**bservation-enhanced **A**lignment **L**earning), a framework designed to enrich the representation space and prevent shortcut learning by leveraging external knowledge. First, we address semantic ambiguity via Explicit Semantic Injection (ESI). To enhance discrimination, we utilize a VLM to generate dual-purpose priors comprising visual detections and linguistic captions. To integrate these priors, we propose a Hierarchical Multi-Stream Integration (HMSI) architecture with a progressive interaction strategy: visually, pixel-word contextualization and deformable sampling transform static crops into query-aware representations; semantically, the referring query filters the caption stream to yield informative descriptions. By fusing these streams, we construct a holistic object representation that is both visually grounded and linguistically explicit. Second, we alleviate sparse supervision via Counterfactual Learning (CFL). To prevent shortcut learning induced by limited data, we leverage an LLM to generate counterfactual hard negatives through stochastic attribute perturbation. These high-interference distractors supervise the model to disentangle subtle cues from salient ones, effectively expanding the semantic representation and mitigating feature collapse. The conceptual mechanism of COAL is illustrated in Fig. 2. Starting from the coupled original distribution, ESI enhances semantic separability through VLM-derived discriminative cues, while CFL promotes compositional discrimination through counterfactual supervision. The resulting representation is expected to be both discriminative and structurally comprehensive, which is further supported by our visualization and quantitative experiments.

The contributions of this work can be summarized as: (1) We systematically identify the Sparsity-Discriminability Paradox in RMOT, where the mismatch between fine-grained discrimination demands and sparse semantic supervision induces shortcut learning and weak compositional generalization. (2) We propose COAL, a unified framework that leverages external knowledge to resolve this paradox. Specifically, ESI enriches semantic representations, CFL introduces counterfactual supervision, and HMSI integrates these priors for knowledge-regularized learning. (3) COAL achieves state-of-the-art performance on both Refer-KITTI and Refer-KITTI-V2 benchmarks, surpassing the second-best method by 7.28% HOTA on the highly challenging Refer-KITTI-V2, demonstrating the effectiveness of knowledge regularization beyond isolated structural optimization.

2 Related Work

Referring Multi-Object Tracking (RMOT). Since the establishment of the Refer-KITTI benchmark by TransRMOT [Wu *et al.*, 2023], RMOT research has primarily focused on isolated optimization of cross-modal interactions. To enhance feature granularity, DeepRMOT [He *et al.*, 2024] and HFF-Tracker [Zhao *et al.*, 2025] introduce hierarchical fusion modules, while CGATracker [Zhuang *et al.*, 2025] and MGLT [Chen *et al.*, 2025] utilize graph alignment and query

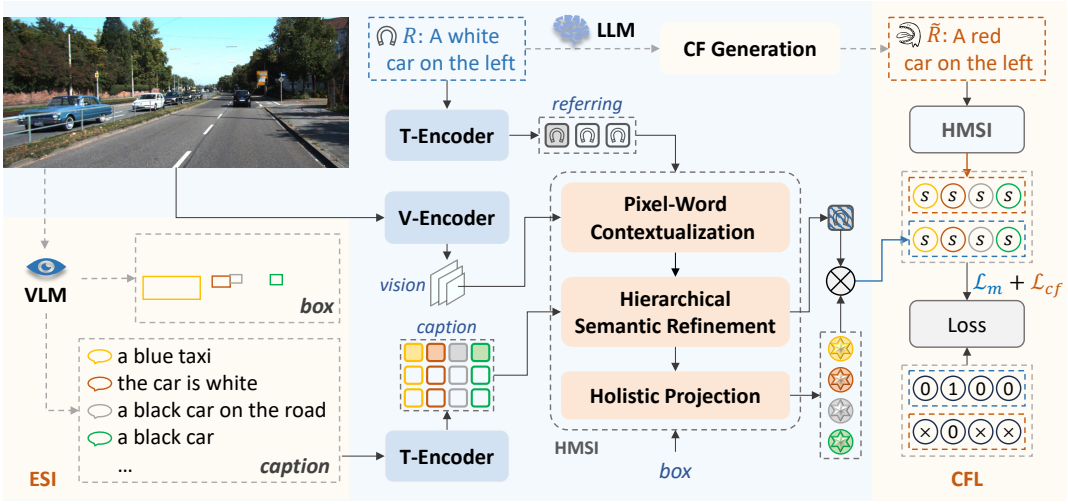

Figure 3: **Overview of the COAL Framework.** It leverages external knowledge through Explicit Semantic Injection (ESI) and Counterfactual Learning (CFL), unified by the Hierarchical Multi-Stream Integration (HMSI) architecture to construct semantically enriched and discriminative object representations. The V-Encoder and T-Encoder refer to the CLIP visual and textual backbones for feature extraction.

bootstrapping to preserve linguistic cues. Other approaches like TellTrack [Huang *et al.*, 2024] and FlexHook [Li *et al.*, 2025a] refine encoder-decoder designs to address data imbalance. iKUN [Du *et al.*, 2024] adopts a plug-and-play strategy with a Neural Kalman Filter. Addressing temporal dynamics, TempRMOT [Zhang *et al.*, 2024] and SK-Track [Li *et al.*, 2025b] incorporate motion modules, while works like LaMOT [Li *et al.*, 2025c], MLS-Track [Ma *et al.*, 2024] and EchoTrack [Lin *et al.*, 2024] expand the task scope to diverse scenarios. CDRMT [Liang *et al.*, 2025] explores cognitive pathways for description disentanglement. Similar explorations on multi-modal perception and spatio-temporal modeling have also been studied in related tracking tasks [Hu *et al.*, 2024; Hu *et al.*, 2025b; Hu *et al.*, 2025a; Hu *et al.*, 2026]. Despite significant breakthroughs, these methods remain confined to a closed data loop. In contrast, we formulate the problem from a data-centric perspective, aiming to mitigate inherent biases in limited data.

Foundation Models in RMOT. Recently, Multi-modal LLMs (MLLMs) have been applied to RMOT, reflecting an emerging trend of leveraging open-world knowledge for domain-specific challenges. For instance, ReferGPT [Chamiti *et al.*, 2025] employs MLLMs for zero-shot tracking via 3D-aware priors, while VMRMOT [Lv *et al.*, 2025] utilizes them to extract motion features for enhanced alignment. In contrast to these approaches, which primarily focus on optimizing multi-modal matching or fusion, we adopt a data-centric perspective to address RMOT's structural contradiction, leveraging external knowledge to regularize the learning process and boost discriminative capacity.

3 Methodology

We propose COAL (Counterfactual and Observation-Enhanced Alignment Learning), a framework designed to resolve the sparsity-discriminability paradox in RMOT via external knowledge regularization. Adopting a Tracking-by-Detection [Zhang *et al.*, 2022] paradigm, COAL focuses on enhancing the semantic discriminability of observations within the tracking pipeline. As illustrated in Fig. 3, the framework is built upon three primary components: (1) **Explicit Semantic Injection** (**ESI**, Sec. 3.1): We first establish a robust observation basis via a VLM, generating dual-purpose priors (boxes and captions) to densify the sparse visual signals. (2) **Hierarchical Multi-Stream Integration** (**HMSI**): To synthesize these priors with visual and linguistic streams, we design the HMSI architecture, integrating the information flow through three progressive stages: Pixel-Word Contextualization (Sec. 3.2) for early implicit grounding, Hierarchical Semantic Refinement (Sec. 3.3) for explicit caption denoising, and Holistic Projection (Sec. 3.4) for unifying multi-modal representations. (3) **Counterfactual Learning** (**CFL**, Sec. 3.5): Driven by an LLM, we generate counterfactual negatives to impose a causal constraint on supervision to rectify shortcut learning.

3.1 Explicit Semantic Injection

To mitigate the ambiguity of implicit visual features supervised by scarce data, we introduce Explicit Semantic Injection (ESI) via a frozen VLM. For each video frame, the VLM executes two parallel tasks: (1) Localization: generating dense object proposals $\mathcal{B}$; and (2) Description: producing descriptive captions $\mathcal{C}$ that materialize subtle visual cues.

This mechanism establishes a dual-purpose observation prior, primarily densifying the semantic representation to serve as the foundational input for the subsequent integration network. Furthermore, by offloading detection and description to the VLM, ESI significantly alleviates the optimization burden, allowing the model to focus on fine-grained multi-modal understanding rather than struggling with basic perception tasks under sparse supervision.

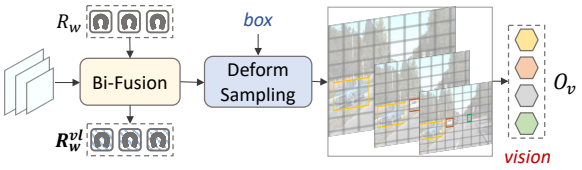

Figure 4: **Illustration of Pixel-Word Contextualization.** It performs bi-directional interaction between visual and referring features, followed by VLM-guided deformable sampling to extract query-aware visual representations.

3.2 Pixel-Word Contextualization

To bridge the modality gap early in the encoding stage, we implement Pixel-Word Contextualization. Given an image and a referring sentence, we extract flattened visual features F_v and tokenized word embeddings $R_w = \{w_1, \ldots, w_L\}$ using frozen CLIP encoders [Radford *et al.*, 2021]. As displayed in Fig. 4, instead of processing them independently, we employ Bi-Fusion [Liu *et al.*, 2024; Chen *et al.*, 2025] to perform bidirectional cross-modal interaction, allowing them to mutually exchange contextual cues:

$$F_v^{vl}, R_w^{vl} = \text{Bi-Fusion}(F_v, R_w). \tag{1}$$

This yields grounded visual representations F_v^{vl} that highlight referred regions, and visually contextualized referring word embeddings R_w^{vl} that drive the subsequent refinement.

Visual Stream. To extract object-centric features, we utilize the high-quality proposals $\mathcal{B} = \{b_i\}_{i=1}^{N}$ generated by the VLM. Guided by these boxes, we reconstruct the grounded representations F_v^{vl} into a multi-scale feature map $\text{MS}(F_v^{vl})$ and apply the standard deformable sampling [Zhu *et al.*, 2020] to dynamically aggregate object-centric features around proposal regions:

$$O_{v,i} = \text{DeformSample}(\text{MS}(F_v^{vl}),\ b_i). \tag{2}$$

Sampled on the linguistically contextualized feature map, O_v encodes both geometric details and linguistic relevance.

3.3 Hierarchical Semantic Refinement

Parallel to the visual stream, we propose this module to refine both referring and caption streams (Fig. 5) through hierarchical filtering and aggregation, ensuring contextually aligned and semantically denoised linguistic representations.

Referring Stream. We synthesize local linguistic cues through the ALG (Aggregation from Local to Global) mechanism. Specifically, the visually-enhanced word features R_w^{vl} are aggregated via cross-attention using the sentence embedding R_s as the query, yielding a referring representation F_r enriched with visual context:

$$F_r = \text{CrossAttn}(Q = R_s, K = R_w, V = R_w^{vl}). \tag{3}$$

Caption Stream. We treat the descriptive captions $\mathcal{C}$ as an explicit semantic stream, tokenized into word and sentence embeddings C_w, C_s, respectively. To filter redundant information, we first propose the FLL (Filtering from Local to Local) mechanism, leveraging the textual alignment between caption and referring words to inject visual context R_w^{vl}:

$$C_w^{vl} = \text{CrossAttn}(Q = C_w, K = R_w, V = R_w^{vl}). \tag{4}$$

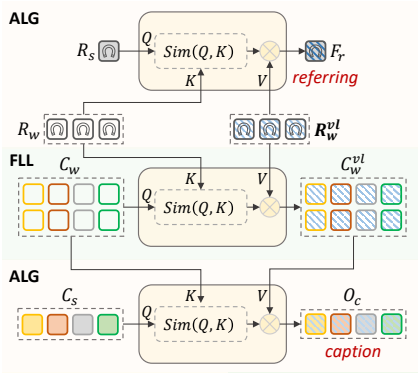

Figure 5: **Illustration of Hierarchical Semantic Refinement.** It refines referring and caption streams via a hierarchical mechanism: Filtering from Local to Local (FLL), to denoise captions using referring cues, and Aggregation from Local to Global (ALG), to synthesize holistic semantic descriptors.

This operation activates caption tokens that resonate with the referring query while enriching them with contextual visual cues. Subsequently, we employ the ALG to synthesize the filtered caption words C_w^{vl} using the sentence embedding C_s:

$$O_c = \text{CrossAttn}(Q = C_s, K = C_w, V = C_w^{vl}). \tag{5}$$

The resulting explicit semantic feature O_c serves as a query-adaptive descriptor, ready for downstream holistic fusion.

3.4 Holistic Projection

To construct a comprehensive representation, we perform a joint projection of the implicit visual stream (O_v), explicit semantic stream (O_c), and spatial geometry. Box coordinates of $\mathcal{B}$ are first encoded into sinusoidal positional embeddings O_p. Treating captions as the semantic counterpart to vision, we construct spatially-aware representations for both streams:

$$\tilde{O}_v = O_v \oplus O_p, \quad \tilde{O}_c = O_c \oplus O_p, \tag{6}$$

where $\tilde{O}_v$ and $\tilde{O}_c$ encapsulate the object's identity within the implicit visual and explicit knowledge spaces, respectively. $\oplus$ represents element-wise addition. To synthesize these complementary views, we project them into a unified latent space:

$$F_o = \mathcal{F}_{\text{fuse}}(\tilde{O}_v \oplus \tilde{O}_c), \tag{7}$$

where $\mathcal{F}_{\text{fuse}}$ denotes a linear projection. This yields a holistic object query F_o that is visually grounded, linguistically explicit, and spatially precise. The final matching score is computed as the cosine similarity between the object query and the referring query: $S = \cos(F_o, F_r)$.

3.5 Counterfactual Learning

To address attribute coupling inherently under semantic sparsity, we introduce Counterfactual Learning (CFL). As illustrated in the right panel of Fig. 3, CFL imposes a causal constraint on supervision to enforce attribute disentanglement.

Counterfactual Generation. For a referring expression R, we harness the reasoning capabilities of an LLM to generate a hard negative query $\tilde{R}$ via stochastic perturbation. The

LLM parses attributes and randomly replaces one (e.g., altering Color while retaining the others). This stochastic strategy functions as an implicit hard negative miner: perturbing a dominant attribute creates an easy negative with minimal loss, whereas perturbing a non-dominant attribute creates a hard negative with high similarity to the target, thus dominating the gradient. This naturally drives the model to focus on discerning subtle, fine-grained attributes beyond salient ones.

Optimization Objective. We formulate the training objective using Binary Cross Entropy (BCE). To stabilize gradient propagation, we linearly map the cosine similarity score $S(u, v) \in [-1, 1]$ to a probability space:

$$P(u, v) = \frac{1}{2}(S(u, v) + 1). \tag{8}$$

The total loss $\mathcal{L} = \mathcal{L}_m + \mathcal{L}_{cf}$ comprises two complementary terms. First, the main grounding loss ($\mathcal{L}_m$) is computed over all N detected objects to minimize the classification error:

$$\mathcal{L}_m = \frac{1}{N}\sum_{i=1}^{N} \ell_i, \quad \ell_i = \begin{cases} -\log(p_i) & \text{if } y_i = 1, \\ -\log(1 - p_i) & \text{if } y_i = 0, \end{cases} \tag{9}$$

where y_i is the ground-truth label, and $p_i = P(F_{o,i}, F_r)$ denotes the matching probability between the i-th object and the global referring query. This term ensures the model correctly identifies the target while suppressing non-target objects.

Second, the counterfactual loss ($\mathcal{L}_{cf}$) is applied exclusively to the target object, which is denoted as o^+. Crucially, since the HMSI network is query-guided, the target's holistic representation dynamically updates from F_{o^+} to $F_{o^+}^{cf}$ when conditioned on the counterfactual query F_r^{cf}. To enforce attribute disentanglement, we require the model to suppress the target under the perturbed description:

$$\mathcal{L}_{cf} = -\log(1 - P(F_{o^+}^{cf},\ F_r^{cf})). \tag{10}$$

Note that we only penalize the original target object o^+ while ignoring other objects, as the perturbed query F_r^{cf} might coincidentally match another object in the scene.

By minimizing the joint objective, our framework establishes a dual-force optimization mechanism. $\mathcal{L}_m$ exerts a *Pull* force for standard alignment. In contrast, $\mathcal{L}_{cf}$ exerts a *Push* force: by teaching the model *what the object is not*, it penalizes spurious correlations against subtle perturbations. This interplay compels the model to acquire compositional understanding, effectively expanding the semantic representation to prevent feature collapse induced by shortcut learning.

4 Experiments

4.1 Datasets and Evaluation Metrics

We systematically evaluate our method on Refer-KITTI [Wu *et al.*, 2023] and Refer-KITTI-V2 [Zhang *et al.*, 2024] to validate its effectiveness under diverse semantic and discriminative conditions. As the most representative and widely adopted benchmarks in RMOT, these two datasets collectively span a comprehensive difficulty spectrum, extending from fundamental scenarios to rigorous high-discriminability challenges. Refer-KITTI established the initial benchmark, comprising 15 training and 3 testing videos with 818 expressions. Refer-KITTI-V2 is a large-scale extension featuring 9,758 diverse expressions across 21 videos. This increased complexity imposes a stringent test on fine-grained discrimination, making it a critical benchmark for validating our framework. We perform ablation studies on both datasets to verify the consistent contribution of the proposed modules.

For evaluation, we adopt HOTA [Luiten *et al.*, 2021] as the primary metric to balance detection and association. We also report DetA and AssA for detailed insights. Moreover, we include auxiliary metrics: DetRe/DetPr to characterize detection performance, AssRe/AssPr to quantify identity consistency, and LocA to assess bounding box quality.

4.2 Implementation Details

We provide comprehensive implementation details, including data preprocessing, network architecture, training strategy, and inference procedure, in the supplementary material. The VLM and LLM components are used offline for proposal/caption generation and counterfactual preparation, respectively, introducing no additional online reasoning overhead during inference.

4.3 Benchmark Results

Tab. 1 presents the comparison with state-of-the-art methods. Our COAL framework achieves the best performance across both benchmarks, validating the effectiveness of integrating external knowledge priors under semantic sparsity.

On the Refer-KITTI benchmark, COAL achieves 53.38% HOTA, outperforming the previous state-of-the-art HFF-Tracker [Zhao *et al.*, 2025] by 0.97%. On the more challenging Refer-KITTI-V2, which features significantly complex semantics and higher discriminability demands, our method demonstrates a substantial advantage. COAL attains 43.46% HOTA, surpassing the second-best method by a margin of 7.28%. The widening performance gap between the two datasets strongly supports our core motivation. Existing methods, focusing on internal structural optimization, tend to saturate on the relatively simpler Refer-KITTI dataset but struggle to generalize to the complex, compositional attributes in Refer-KITTI-V2. We attribute this robustness to the effective utilization of external knowledge through three key components: (1) Explicit knowledge injection (via VLM) densifies the observation space, enabling the recognition of subtle attributes that implicit features often miss; (2) Counterfactual alignment (via LLM) enforces attribute disentanglement, preventing shortcut learning in sparse semantic scenarios; and (3) Hierarchical integration (via HMSI) effectively synthesizes these diverse priors into a coherent representation. This demonstrates that knowledge-regularized integration is a highly effective strategy for resolving the sparsity-discriminability paradox, providing a promising direction beyond isolated structural optimization for RMOT.

4.4 Ablation Studies

We conduct extensive ablation studies on both Refer-KITTI and Refer-KITTI-V2 benchmarks to dissect the effectiveness of our framework. The analysis is structured around two fundamental questions. (1) *What to Inject:* The necessity of the

Method	Publication	HOTA↑	DetA↑	AssA↑	DetRe↑	DetPr↑	AssRe↑	AssPr↑	LocA↑
Refer-KITTI									
TransRMOT [Wu *et al.*, 2023]	CVPR 2023	46.56	37.97	57.33	49.69	60.10	60.02	89.67	90.33
DeepRMOT [He *et al.*, 2024]	ICASSP 2024	39.55	30.12	53.23	41.91	47.47	58.47	82.16	80.49
iKUN [Du *et al.*, 2024]	CVPR 2024	48.84	35.74	66.80	51.97	52.25	72.95	87.09	–
EchoTrack [Lin *et al.*, 2024]	ITS 2024	48.86	41.26	57.59	53.42	62.83	61.61	89.33	90.74
TempRMOT [Zhang *et al.*, 2024]	arXiv 2024	52.21	40.95	66.75	55.65	59.25	71.82	87.76	90.40
CGATracker [Zhuang *et al.*, 2025]	TCSVT 2025	48.81	38.66	61.77	51.67	59.12	65.87	**90.26**	–
MGLT [Chen *et al.*, 2025]	TIM 2025	49.25	37.09	65.50	–	–	–	–	–
CDRMT [Liang *et al.*, 2025]	IF 2025	49.35	40.34	60.56	54.54	59.30	64.70	89.80	90.61
SKTrack [Li *et al.*, 2025b]	TMM 2025	50.85	**42.24**	61.39	58.10	59.24	66.68	88.76	90.43
HFF-Tracker [Zhao *et al.*, 2025]	AAAI 2025	52.41	41.29	66.65	53.42	**62.89**	71.48	88.96	90.76
COAL	**Ours**	**53.38**	40.89	**69.74**	**60.32**	54.73	**76.25**	86.68	**90.99**
Refer-KITTI-V2									
TransRMOT [Wu *et al.*, 2023]	CVPR 2023	31.00	19.40	49.68	36.41	28.97	54.59	82.29	89.82
iKUN [Du *et al.*, 2024]	CVPR 2024	10.32	2.17	49.77	2.36	19.75	58.48	68.64	74.56
TempRMOT [Zhang *et al.*, 2024]	arXiv 2024	35.04	22.97	53.58	34.23	40.41	59.50	81.29	90.07
CDRMT [Liang *et al.*, 2025]	IF 2025	31.99	20.37	50.35	26.40	**46.26**	53.40	**85.90**	90.36
CGATracker [Zhuang *et al.*, 2025]	TCSVT 2025	33.19	22.04	50.13	38.00	33.88	54.73	84.90	–
SKTrack [Li *et al.*, 2025b]	TMM 2025	35.29	23.87	52.35	39.97	36.48	57.45	84.23	88.89
HFF-Tracker [Zhao *et al.*, 2025]	AAAI 2025	36.18	24.64	53.27	36.86	41.83	59.42	81.40	89.77
COAL	**Ours**	**43.46**	**32.83**	**57.62**	**60.20**	41.19	**66.23**	80.33	**90.38**

Table 1: **Comparison with State-of-the-Art Methods on the Refer-KITTI and Refer-KITTI-V2 Benchmarks.** The best results for each metric are highlighted in **bold**.

Knowledge Priors		Refer-KITTI			Refer-KITTI-V2		
ESI	CFL	HOTA↑	DetA↑	AssA↑	HOTA↑	DetA↑	AssA↑
✗	✗	46.84	32.87	66.76	37.29	25.12	55.44
✓	✗	48.64	35.36	66.95	39.38	27.66	56.15
✗	✓	51.48	39.15	67.75	40.01	28.43	56.40
✓	✓	53.38	40.89	69.74	43.46	32.83	57.62

Table 2: **Ablation Study on the Impact of External Knowledge Integration.** We validate the contribution of Explicit Semantic Injection (ESI) and Counterfactual Learning (CFL) on both Refer-KITTI and Refer-KITTI-V2 benchmarks.

introduced knowledge priors; and (2) *How to inject:* The rationality of the proposed hierarchical integration architecture.

Impact of External Knowledge Integration. To investigate the contributions of the two external knowledge integration strategies, we define two ablation settings: (1) w/o ESI: We remove the ESI, disabling the filtering and aggregation of captions in the HMSI module and relying solely on the implicit visual stream. (2) w/o CFL: We deactivate the CFL branch and remove the related loss $\mathcal{L}_{cf}$ during training. As shown in Tab. 2, ESI consistently improves performance over the baseline. For instance, on the more challenging Refer-KITTI-V2 dataset, HOTA improves from 37.29% to 39.38%. This confirms that VLM captions serve as effective semantic anchors to mitigate visual ambiguity. Introducing CFL yields even more significant gains, boosting HOTA to 40.01%. This demonstrates its crucial role in preventing shortcut learning. Ultimately, the full model achieves the highest performance, validating that representation enrichment (via ESI) and supervision augmentation (via CFL) operate as complementary mechanisms to address the structural sparsity-discriminability paradox in RMOT.

Efficacy of Fusion Architecture. To validate our hierarchical interaction design, we compare the full HMSI architecture against simplified variants in Tab. 3. First, we investigate the impact of early pixel-word interaction by removing the Bi-Fusion module (Sec. 3.2). The significant performance drop from 41.42% to 38.43% HOTA on Refer-KITTI-V2 confirms that the mutual injection is critical for subsequent refinement: it not only generates query-aware visual representations but also enriches referring signals with visual context. Second, we examine the necessity of caption refinement (Sec. 3.3). Substituting the referring-guided refinement O_c with raw sentence-level captions C_s leads to suboptimal results. The superiority of the full model validates that using the referring query as a semantic filter to denoise the caption stream is essential for extracting discriminative cues. Interestingly, we observe a compensation effect between the two modules: The performance gain of Bi-Fusion narrows from 2.99% to 0.61% when caption refinement is active. This suggests that the refined caption stream can partially compensate for the lack of early cross-modal contextualization by providing robust external priors. Nevertheless, optimal performance still requires the synergy of both components.

Together, these experiments demonstrate that effective

Fusion Architecture		Refer-KITTI			Refer-KITTI-V2		
Bi-Fusion	Caption Refine	HOTA↑	DetA↑	AssA↑	HOTA↑	DetA↑	AssA↑
✗	✗	49.34	35.27	69.07	38.43	27.16	54.47
✓	✗	52.39	39.88	68.87	41.42	30.81	55.76
✗	✓	52.85	41.04	68.10	42.85	32.74	56.19
✓	✓	53.38	40.89	69.74	43.46	32.83	57.62

Table 3: **Ablation Study on Hierarchical Multi-Stream Integration (HMSI).** We analyze the necessity of key interaction mechanisms: Bi-Fusion (the core of Pixel-Word Contextualization, Sec. 3.2) and Caption Refinement (the core of Hierarchical Semantic Refinement, Sec. 3.3), on both Refer-KITTI and Refer-KITTI-V2 benchmarks.

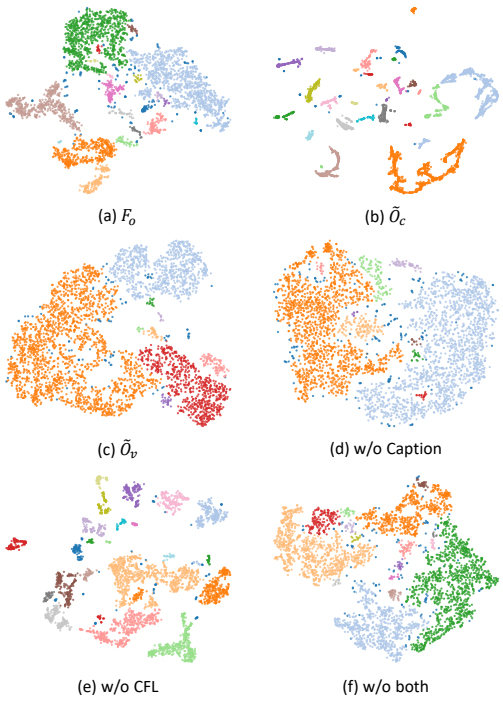

Figure 6: **Visualization of t-SNE Embeddings for Object Representations.** Subplots (a)-(c) illustrate features from our full model: (a) the holistic representation F_o (Equation 7), (b) the explicit caption component $\widetilde{O}_c$ (Equation 6), and (c) the implicit visual component $\widetilde{O}_v$. (d)-(f) compare holistic representations from ablation variants: (d) w/o Caption, (e) w/o CFL, and (f) Baseline (w/o both). Colors indicate different semantic clusters.

knowledge regularization requires both informative priors (*what*) and hierarchical integration (*how*).

4.5 Visualization

We visualize t-SNE embeddings of object representations from a validation sequence to qualitatively elucidate how Explicit Semantic Injection (ESI) and Counterfactual Learning (CFL) reshape the semantic manifold. By mapping the ablation settings from Tab. 2 to visual clusters, this analysis provides an intuitive understanding of the quantitative gains.

The Anchoring Effect of ESI. Comparing the implicit visual stream (Fig. 6c) with the explicit caption stream (Fig. 6b) reveals a contrast: while visual features overlap severely due to homogeneity, captions form compact clusters, serving as discriminative semantic anchors. Integrating these anchors (Fig. 6a) alleviates the ambiguity present in the baseline (Fig. 6f). This refinement persists even without CFL (comparing Fig. 6e vs. 6f). Moreover, we observe a phenomenon of implicit knowledge distillation: the visual component of our full model (Fig. 6c) shows better separability than the caption-free model (Fig. 6d). This indicates that the caption stream acts as a reference, guiding the model to discern discriminative visual cues even within the implicit branch.

The Disentangling Effect of CFL. Comparing the model without CFL (Fig. 6e) to the full model (Fig. 6a) elucidates the mechanism for mitigating semantic collapse. The former forms extremely tight, point-like clusters. Although superficially appealing, this is symptomatic of feature collapse by shortcut learning. In contrast, introducing CFL forces these clusters to expand and fill the semantic space (Fig. 6a). This dispersion is theoretically desirable: by distinguishing counterfactual hard negatives, CFL prevents the representation from degenerating into a single-attribute dominated embedding. Instead, it enforces the encoding of fine-grained nuances, resulting in a disentangled representation that faithfully reflects the intrinsic diversity of the data manifold.

In essence, the two modules operate in a complementary manner: ESI acts as an anchor, injecting distinct semantics into ambiguous visual features to sharpen semantic boundaries; meanwhile, CFL promotes expansion within these clusters, preventing feature collapse into single-dimensional shortcuts. This synergy ensures a representation that is both semantically discriminative and structurally comprehensive.

5 Conclusion

In this work, we identify and analyze the Sparsity–Discriminability Paradox in RMOT, where fine-grained discrimination requirements conflict with sparse semantic supervision. To address this challenge, we propose COAL, a knowledge-centric alignment framework that incorporates external semantic regularization beyond isolated structural optimization. By integrating Explicit Semantic Injection (ESI) and Counterfactual Learning (CFL) within a Hierarchical Multi-Stream Integration (HMSI) architecture, COAL alleviates shortcut learning and improves compositional discrimination under scarce supervision. Extensive experiments on Refer-KITTI and Refer-KITTI-V2 demonstrate the effectiveness of systematically introducing external semantic knowledge from foundation models for mitigating discrimination bottlenecks in small-scale RMOT tasks.

Acknowledgments

We are grateful for the support of the National Natural Science Foundation of China (No. 62271143), the Frontier Technologies R&D Program of Jiangsu (No. BF2024060), and the Big Data Computing Center of Southeast University. We also thank the IDEA Research Team for providing academic access to the DINO-X API.

References

[Chamiti *et al.*, 2025] Tzoulio Chamiti, Leandro Di Bella, Adrian Munteanu, and Nikos Deligiannis. Refergpt: Towards zero-shot referring multi-object tracking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3849–3858, 2025.

[Chen *et al.*, 2025] Jiajun Chen, Jiacheng Lin, Guojin Zhong, You Yao, and Zhiyong Li. Multi-granularity localization transformer with collaborative understanding for referring multi-object tracking. *IEEE Transactions on Instrumentation and Measurement*, 2025.

[Du *et al.*, 2024] Yunhao Du, Cheng Lei, Zhicheng Zhao, and Fei Su. ikun: Speak to trackers without retraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19135–19144, 2024.

[He *et al.*, 2024] Wenyan He, Yajun Jian, Yang Lu, and Hanzi Wang. Visual-linguistic representation learning with deep cross-modality fusion for referring multi-object tracking. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6310–6314. IEEE, 2024.

[Hu *et al.*, 2024] Xiantao Hu, Bineng Zhong, Qihua Liang, Shengping Zhang, Ning Li, and Xianxian Li. Toward modalities correlation for rgb-t tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10):9102–9111, 2024.

[Hu *et al.*, 2025a] Xiantao Hu, Ying Tai, Xu Zhao, Chen Zhao, Zhenyu Zhang, Jun Li, Bineng Zhong, and Jian Yang. Exploiting multimodal spatial-temporal patterns for video object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3581–3589, 2025.

[Hu *et al.*, 2025b] Xiantao Hu, Bineng Zhong, Qihua Liang, Liangtao Shi, Zhiyi Mo, Ying Tai, and Jian Yang. Adaptive perception for unified visual multi-modal object tracking. *IEEE Transactions on Artificial Intelligence*, 2025.

[Hu *et al.*, 2026] Xiantao Hu, Fansheng Zeng, Bineng Zhong, Zhangyong Tang, Wenxuan Fang, Jun Li, Ying Tai, and Jian Yang. Curriculum adaptation for one-stream rgb–t tracking. *Pattern Recognition*, page 113494, 2026.

[Huang *et al.*, 2024] Wenjun Huang, Yang Ni, Hanning Chen, Yirui He, Ian Bryant, Yezi Liu, and Mohsen Imani. Tell me what to track: Infusing robust language guidance for enhanced referring multi-object tracking. *arXiv preprint arXiv:2412.12561*, 2024.

[Li *et al.*, 2025a] Weize Li, Yunhao Du, Qixiang Yin, Zhicheng Zhao, Fei Su, and Daqi Liu. Just functioning as a hook for two-stage referring multi-object tracking. *arXiv preprint arXiv:2503.07516*, 2025.

[Li *et al.*, 2025b] Yizhe Li, Sanping Zhou, Zheng Qin, and Le Wang. Visual-linguistic feature alignment with semantic and kinematic guidance for referring multi-object tracking. *IEEE Transactions on Multimedia*, 2025.

[Li *et al.*, 2025c] Yunhao Li, Xiaoqiong Liu, Luke Liu, Heng Fan, and Libo Zhang. Lamot: Language-guided multi-object tracking. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6816–6822. IEEE, 2025.

[Liang *et al.*, 2025] Shaofeng Liang, Runwei Guan, Wangwang Lian, Daizong Liu, Xiaolou Sun, Dongming Wu, Yutao Yue, Weiping Ding, and Hui Xiong. Cognitive disentanglement for referring multi-object tracking. *Information Fusion*, page 103349, 2025.

[Lin *et al.*, 2024] Jiacheng Lin, Jiajun Chen, Kunyu Peng, Xuan He, Zhiyong Li, Rainer Stiefelhagen, and Kailun Yang. Echotrack: Auditory referring multi-object tracking for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[Liu *et al.*, 2023] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23592–23601, 2023.

[Liu *et al.*, 2024] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024.

[Luiten *et al.*, 2021] 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.

[Lv *et al.*, 2025] Weiyi Lv, Ning Zhang, Hanyang Sun, Haoran Jiang, Kai Zhao, Jing Xiao, and Dan Zeng. Vision-motion-reference alignment for referring multi-object tracking via multi-modal large language models. *arXiv preprint arXiv:2511.17681*, 2025.

[Ma *et al.*, 2024] Zeliang Ma, Song Yang, Zhe Cui, Zhicheng Zhao, Fei Su, Delong Liu, and Jingyu Wang. Mls-track: Multilevel semantic interaction in rmot. *arXiv preprint arXiv:2404.12031*, 2024.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[Wu *et al.*, 2023] Dongming Wu, Wencheng Han, Tiancai Wang, Xingping Dong, Xiangyu Zhang, and Jianbing Shen. Referring multi-object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14633–14642, 2023.

[Zhang *et al.*, 2022] 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*, pages 1–21. Springer, 2022.

[Zhang *et al.*, 2024] Yani Zhang, Dongming Wu, Wencheng Han, and Xingping Dong. Bootstrapping referring multi-object tracking. *arXiv preprint arXiv:2406.05039*, 2024.

[Zhao *et al.*, 2025] Zeyong Zhao, Yanchao Hao, Minghao Zhang, Qingbin Liu, Bo Li, Dianbo Sui, Shizhu He, and Xi Chen. Hff-tracker: A hierarchical fine-grained fusion tracker for referring multi-object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 10528–10536, 2025.

[Zhu *et al.*, 2020] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

[Zhuang *et al.*, 2025] Siping Zhuang, Guangyao Li, Qiangqiang Wu, Yang Lu, Hai-Miao Hu, and Hanzi Wang. Cgatracker: Correlation-aware graph alignment for referring multi-object tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.