CD-RMOT-BENCH: BENCHMARKING CROSS-DOMAIN REFERRING MULTI-OBJECT TRACKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Referring multi-object tracking (RMOT) extends tracking from category-driven perception to language-guided understanding by grounding object trajectories in natural-language expressions. Despite recent progress, existing RMOT studies are largely conducted under in-domain settings, leaving the robustness of language-conditioned tracking under inevitable visual domain shifts unexplored. In this paper, we study **Cross-Domain Referring Multi-Object Tracking (CD-RMOT)**, a new and challenging problem that evaluates whether an RMOT model trained on a labeled source domain can reliably follow natural-language expressions in an unlabeled target domain with different visual conditions. To support systematic study, we construct **CD-RMOT-Bench**, a unified benchmark that combines real clear-domain referring tracking data, aligned digital-twin variants, and real adverse-domain videos. CD-RMOT-Bench enables both controlled weather/viewpoint shift analysis and realistic synthetic–real transfer evaluation under a shared RMOT protocol. Further, we provide a **Query-Centric Adaptation (QCA)** framework for unsupervised domain adaptation (UDA), designed to stabilize the query space that bridges visual trajectories and referring expressions. Extensive experiments reveal that domain shifts severely degrade RMOT performance, affecting both detection and association, with qualitative examples also showing expression-inconsistent target selection. QCA establishes a strong baseline, while CD-RMOT-Bench opens a new direction for robust language-guided tracking across visual domains.

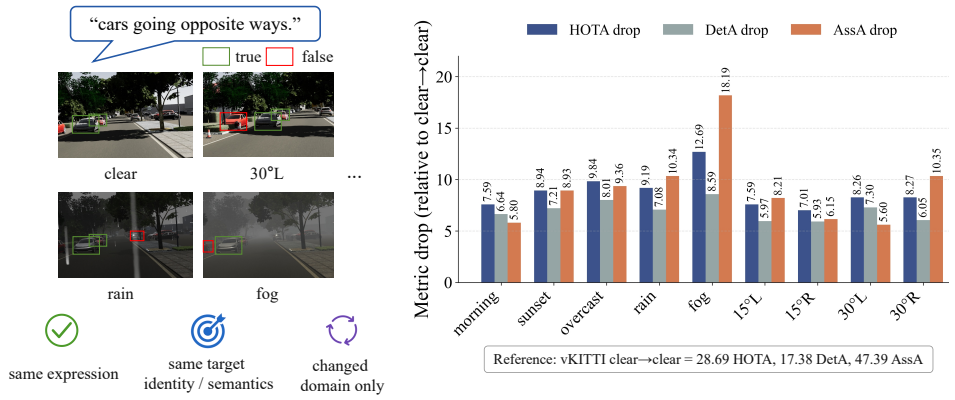
1 INTRODUCTION

Multi-Object Tracking (MOT) is undergoing a paradigm shift from passive, category-based perception towards active, language-guided understanding, formalized as Referring Multi-Object Tracking. Unlike traditional MOT, RMOT aims to simultaneously locate and track specific targets in video sequences based on natural language expressions, and has shown potential in human-computer interaction, Embodied AI, autonomous driving, *etc.*, where intelligent agents must comprehend and track specific targets driven by complex human intentions in highly dynamic environments.

Recent RMOT benchmarks and methods have made encouraging progress under the standard in-domain setting, where training and testing videos are usually drawn from the same visual distribution. However, when deployed in real-world scenarios, perception systems inevitably encounter severe domain shifts, including extreme weather variations, unpredictable sensor viewpoint changes, and inherent synthetic-to-real realism gaps. As shown in Fig. 1, such shifts pose a critical challenge to existing RMOT models. Fig. 1(a) presents a controlled case where the referring expression and target semantics remain unchanged, while only the visual domain varies across weather and viewpoint conditions. Although the model can still detect visually plausible objects, its selected trajectories may deviate from the language-specified targets. Fig. 1(b) reports aggregate tracking metrics, showing performance drops relative to the in-domain reference. These observations indicate that current RMOT evaluation largely overlooks a fundamental question: *can a language-conditioned tracker remain reliable when the visual domain changes*?

To address this gap, we introduce a novel and challenging problem, Cross-Domain Referring Multi-Object Tracking, which evaluates whether an RMOT model trained in a labeled source domain can transfer to an unlabeled target domain with different visual conditions. The core challenge of CD-RMOT lies in maintaining the semantic fidelity of language-guided tracking trajectories

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.



(a) Controlled domain-shift setting for RMOT

(b) Performance drop relative to vKITTI clear→clear reference

Figure 1: Controlled domain shift in RMOT. (a) Under the same referring expression and target semantics, changing only the visual domain can cause the tracker to select expression-inconsistent trajectories. (b) Existing RMOT model Zhang et al. (2024) shows clear HOTA, DetA, and AssA drops across weather and viewpoint shifts relative to the in-domain *vKITTI clear→clear* reference.

when the underlying visual distribution shifts dramatically, without access to target-domain tracking annotations.

A principled benchmark for this new problem is difficult to build. Real-world videos naturally contain domain variations, but they also change scene layout, object population, camera motion, and target interactions, making it hard to determine whether failure comes from domain shift or from different semantic content. Conversely, purely synthetic benchmarks can provide controllable variations, but they may not reflect the realism gap faced by practical tracking systems. What is needed is a benchmark that *combines controlled domain shift with realistic cross-domain transfer*, so that models can be evaluated along both axes. To this end, we introduce CD-RMOT-Bench, a unified benchmark for CD-RMOT, which integrates real-world referring annotations, aligned digital-twin variants, and adverse real-domain videos into a single evaluation suite. Its controlled part enables weather and viewpoint changes to be studied while preserving the underlying scene structure and referring semantics, while its real-domain part evaluates transfer across synthetic–real domains.

Further, we propose a Query-Centric Adaptation framework for UDA in CD-RMOT. The method is motivated by the need to maintain language-conditioned temporal queries across domains, as these queries connect visual representations, trajectory state, and target selection. To address this issue, we introduce temporal teacher-student consistency to stabilize target-domain queries, adaptive domain adversarial learning to align transferable query representations, and language-guided adaptation to preserve the semantic correspondence between queries and referring expressions. Extensive experiments on CD-RMOT-Bench show that visual domain shifts degrade both detection and association performance. Qualitative examples further illustrate expression-inconsistent target selection. QCA improves tracking across weather, viewpoint, and synthetic–real transfer settings, establishing a strong baseline for further study.

The main contributions are summarized as follows:

- We study the problem of **Cross-Domain Referring Multi-Object Tracking**, highlighting the critical need for robust visual-language tracking under domain shifts.
- We construct **CD-RMOT-Bench**, a unified CD-RMOT benchmark covering controlled weather and viewpoint variations and realistic synthetic–real domain transfer under a shared evaluation protocol. Its controlled core, Refer-vKITTI, is built through per-scene identity mapping, domain-specific expression-track transfer, and residual quality control.
- We provide a query-centric adaptation method that improves CD-RMOT through reliable query stabilization, task-relevant query alignment, and language-anchored query calibration. Extensive experiments demonstrate the severe limitations of existing methods under domain shifts and validate the effectiveness of the proposed framework as a strong baseline for future research.

2 Related Work

Referring Multi-Object Tracking. Language-guided tracking selects targets through natural-language expressions. Modern MOT has developed detection, association, and temporal reasoning mechanisms, from online tracking-by-detection methods (Bewley et al., 2016; Wojke et al., 2017; Bergmann et al., 2019; Zhou et al., 2020; Zhang et al., 2021; 2022; Cao et al., 2023) to transformer/query-based trackers (Sun et al., 2020; Meinhardt et al., 2022; Zeng et al., 2022; Zhang et al., 2023; Yan et al., 2022). RMOT adds open-ended expression grounding to these tracking foundations.

Referring multi-object tracking requires the model to track only the objects specified by a natural-language query. TransRMOT (Wu et al., 2023a) establishes the RMOT task and benchmark, and TempRMOT (Zhang et al., 2024) expands linguistic difficulty through Refer-KITTI-V2. Recent RMOT studies further improve language fusion, plug-and-play referring control, multimodal cues, fine-grained semantic guidance, and reasoning-oriented target selection, including iKUN (Du et al., 2024), EchoTrack (Lin et al., 2024), DeepRMOT (He et al., 2024), MGLT (Chen et al., 2025a), DKGTrack (Li et al., 2025), and recent reasoning or zero-shot variants (Chen et al., 2025b; Chamiti et al., 2025). These works mainly evaluate closed or nearly in-domain behavior, leaving cross-domain robustness of expression-conditioned trajectories underexplored.

A related line is referring video object segmentation (RVOS), where language guides temporal mask prediction. Representative methods include MTTR (Botach et al., 2022), ReferFormer (Wu et al., 2022), OnlineRefer (Wu et al., 2023b), MUTR (Yan et al., 2024), and ReferDINO (Liang et al., 2025). We adapt MUTR and ReferDINO through a mask-to-box interface as cross-task references; RMOT evaluates identity-aware trajectories and association under shift.

Cross-Domain Adaptation. Cross-domain adaptation has been widely studied for recognition, detection, and structured prediction. General adaptation methods reduce distribution mismatch through discrepancy minimization or adversarial learning, including DAN (Long et al., 2015), DANN (Ganin et al., 2016), ADDA (Tzeng et al., 2017), CDAN (Long et al., 2018), and MCD (Saito et al., 2018). For object detection, DA-Faster R-CNN (Chen et al., 2018), SWDA (Saito et al., 2019), MTOR (Cai et al., 2019), Unbiased Mean Teacher (Deng et al., 2021), Adaptive Teacher (Li et al., 2022), Probabilistic Teacher (Chen et al., 2022), and recent prompt-based adaptation methods (Zhang et al., 2025b) show that instance-level alignment and teacher–student training can improve transfer under unlabeled target domains.

Video and test-time adaptation address temporal and non-stationary shifts. TA^{3}N (Chen et al., 2019) studies temporal attentive alignment; Tent (Wang et al., 2021), CoTTA (Wang et al., 2022), EATA (Niu et al., 2022), PETAL (Brahma & Rai, 2023), and synthetic-to-real video studies (Zhang et al., 2025a; Wang et al., 2025) extend adaptation beyond static images. These works motivate target-query consistency, alignment, and language-guided correction. CD-RMOT-Bench evaluates them through expression-conditioned multi-object trajectories.

Benchmarks for Visual Domain Shift. Driving and tracking benchmarks such as KITTI (Geiger et al., 2012), MOTChallenge (Milan et al., 2016), Cityscapes (Cordts et al., 2016), BDD100K (Yu et al., 2020), and nuScenes (Caesar et al., 2020) have shaped progress in perception, but they do not provide controlled variants where scene semantics and referring expressions remain fixed while the visual domain changes. Synthetic and digital-twin datasets, including Virtual KITTI (Gaidon et al., 2016), Virtual KITTI 2 (Cabon et al., 2020), SYNTHIA (Ros et al., 2016), GTA5 (Richter et al., 2016), and SHIFT (Sun et al., 2022), support structured domain variation, but are not designed for referring-expression-conditioned tracking. General domain-shift benchmarks such as VisDA (Peng et al., 2017), DomainNet (Peng et al., 2019), and WILDS (Koh et al., 2021) similarly do not evaluate language-conditioned multi-object trajectories.

Existing RMOT datasets provide language-conditioned trajectories, while controlled synthetic datasets provide domain variation. Recent extensions broaden evaluation toward reasoning-intensive and aerial scenarios (Chen et al., 2025b; 2026). CD-RMOT-Bench adds paired driving-domain shifts by transferring Refer-KITTI-V2 expressions to aligned Virtual KITTI 2 scenes and combining them with real-domain anchors.

3 CD-RMOT-BENCH OVERVIEW

3.1 MOTIVATION AND DESIGN

CD-RMOT-Bench evaluates unsupervised domain adaptation for RMOT using target adaptation videos and referring expressions, but without their tracking annotations. Source-trained transfer is retained as a no-adaptation reference, and adaptation and evaluation use disjoint scenes (Appendix A.2).

The benchmark has two axes: **controlled domain shift** with aligned weather/viewpoint variants and **synthetic–real transfer** across datasets. The paired variants constrain scene-content variation; real-domain transfer also changes scene content and image formation.

Refer-KITTI-V2 (Zhang et al., 2024) supplies real clear-domain expressions and trajectories. Refer-vKITTI transfers them to aligned Virtual KITTI 2 (Cabon et al., 2020) weather/viewpoint variants; Refer-BDD (Lin et al., 2024) supplies real adverse-domain evaluation. All three use the same RMOT output format.

3.2 BENCHMARK CONSTRUCTION

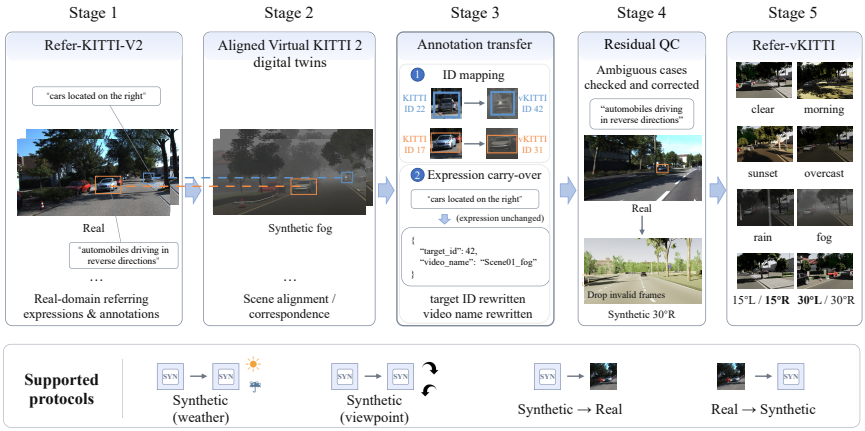


Figure 2: Refer-vKITTI construction and evaluation. Scene-local identity mappings transfer referring trajectories from Refer-KITTI-V2 to aligned Virtual KITTI 2 domains; visibility filtering checks the rendered targets. Refer-BDD adds real-domain transfer.

Fig. 2 summarizes the mapping-then-transfer pipeline. We pair five Refer-KITTI-V2 scenes with their Virtual KITTI 2 digital twins, collect frame-level spatial evidence, and resolve scene-local identity correspondences using trajectory geometry and temporal overlap. Expressions are preserved while target identities and video names are mapped into ten weather/viewpoint domains; visibility filtering removes frame targets absent from the rendered scene. Refer-BDD supplies the real adverse-domain anchor through the shared expression-conditioned trajectory format. Appendix A gives the matching score and transfer rules.

Construction quality. The pipeline retains 272,043 of 285,160 transferred expression–track pairs (95.40%) after visibility filtering. Of 6,887 temporally overlapping identity-pair hypotheses, 374 have IoU evidence and 251 enter the final scene-local one-to-one mapping. Two independent auditors agreed on 369 of the 374 supported candidates before adjudication ($\kappa = 0.970$); a stratified review checked 300 transferred expression–track instances. Appendix A.1 reports these audits and distinct counting units.

3.3 STATISTICS AND EVALUATION PROTOCOL

Table 1 reports the scale, role, and domain coverage of each subset and the overall benchmark.

Scale and expression cues. Refer-vKITTI spans six weather/illumination conditions and four camera shifts across five aligned scenes.

Table 1: Overview of CD-RMOT-Bench. Refer-KITTI-V2, Refer-vKITTI, and Refer-BDD respectively serve as the real clear-domain anchor, controlled digital-twin core, and real adverse-domain anchor for unified CD-RMOT evaluation.

Dataset	Type	Videos	Frames	Exprs.	Expr./Video	Words/Expr.	Coverage
Refer-KITTI-V2 (Zhang et al., 2024)	Real	21	8,008	9,778	465.62	6.64	Real clear scenes
Refer-vKITTI	Synthetic	50	20,560	27,730	554.60	7.26	10 controlled domains
Refer-BDD (Lin et al., 2024)	Real	50	10,005	4,610	92.20	6.30	Real adverse scenes
CD-RMOT-Bench	Mixed	121	38,573	42,118	348.08	6.91	Weather/viewpoint; synthetic–real transfer

Appendix Fig. 5 visualizes dataset composition and four non-exclusive expression-cue groups: appearance, spatial, motion, and relational descriptions. These coarse keyword counts characterize language coverage.

Evaluation protocols. Weather and viewpoint protocols transfer from vKITTI clear to fog, rain, or shifted camera views. Synthetic-to-real protocols target Refer-KITTI-V2 and Refer-BDD; the reverse protocol transfers from Refer-KITTI-V2 to vKITTI fog. Appendix A.2 specifies the scene-disjoint splits and data-access rules.

In controlled transfer, source training and target adaptation use the aligned Scene06/18/20 variants; evaluation uses Scene01/02. Synthetic-to-real evaluation complements these paired visual changes with differences in scene content and image formation.

4 Query-Centric Adaptation

4.1 UDA Formulation and Optimization

QCA is a UDA reference method built on TempRMOT (Zhang et al., 2024). Let $\mathcal{D}_s = \{(V_s, E_s, Y_s)\}$ contain labeled source videos, expressions, and trajectories, and $\mathcal{D}_t^{\text{adapt}} = \{(V_t, E_t)\}$ contain target adaptation videos and expressions. Source trajectories Y_s supervise tracking, while an EMA teacher produces target pseudo boxes for consistency learning.

The student f_θ outputs temporal query features $q_i \in \mathbb{R}^d$, normalized boxes $b_i \in [0,1]^4$, objectness $u_i = \sigma(o_i)$, and referring confidence $v_i = \sigma(r_i)$. The teacher $f_{\bar{\theta}}$ has the same tracker architecture and is evaluated without gradients. Fig. 3 summarizes the three components. With sg denoting stop-gradient, the training objective is

$$\mathcal{L} = \mathcal{L}^s_{\text{RMOT}} + 0.1\mathcal{L}^t_{\text{MT}} + 0.1\mathcal{L}^{s,t}_{\text{ACDA}} + 0.1\mathcal{L}^s_{\text{LGCA}}. \tag{1}$$

Source supervision, target consistency, and domain alignment update the student; LGCA adds language-conditioned query regularization. After each optimizer step, the teacher parameters update as $\bar{\theta} \leftarrow 0.999\bar{\theta} + 0.001\theta$. The teacher and domain discriminator are training-time components.

4.2 Query Objectives and Inference

CDT-MT: target consistency. Cross-Domain Temporal Mean Teacher ranks target queries by teacher reliability $\rho_i = u_i^{\text{tea}} v_i^{\text{tea}}$ and cross-model agreement a_i. The selection score is $\psi_i = 0.65\rho_i + 0.35a_i$. For the selected set Ω_t, detached weights $w_i = \mathrm{sg}(\psi_i)/(\sum_{j\in\Omega_t} \mathrm{sg}(\psi_j) + \varepsilon)$ give

$$\mathcal{L}^t_{\text{MT}} = \eta_e \sum_{i\in\Omega_t} w_i \left[\text{MSE}(q_i^{\text{stu}}, \mathrm{sg}(q_i^{\text{tea}})) + \tfrac{1}{2}\,\text{SL1}(b_i^{\text{stu}}, \mathrm{sg}(b_i^{\text{tea}})) + \tfrac{1}{2}(v_i^{\text{stu}} - \mathrm{sg}(v_i^{\text{tea}}))^2\right]. \tag{2}$$

MSE and smooth-L_1 (SL1) are coordinate means; $\eta_e = \min(1, (e+1)/5)$ for zero-based epoch e. Appendix B.1 specifies final-frame slot selection and teacher-pseudo-box matching.

ACDA: task-relevant domain alignment. Adaptive Curricular Domain Adversarial learning selects up to 16 queries per source or target clip by $c_i = u_i v_i$. For their union $\mathcal{Q}$, a domain classifier g_ϕ predicts the dataset domain label d_i from query z_i:

$$\mathcal{L}^{s,t}_{\text{ACDA}} = \frac{1}{|\mathcal{Q}|} \sum_{i\in\mathcal{Q}} \mathrm{sg}(\omega_i)\,\text{CE}(g_\phi(\text{GRL}_{\lambda_e}(z_i)), d_i)\,. \tag{3}$$

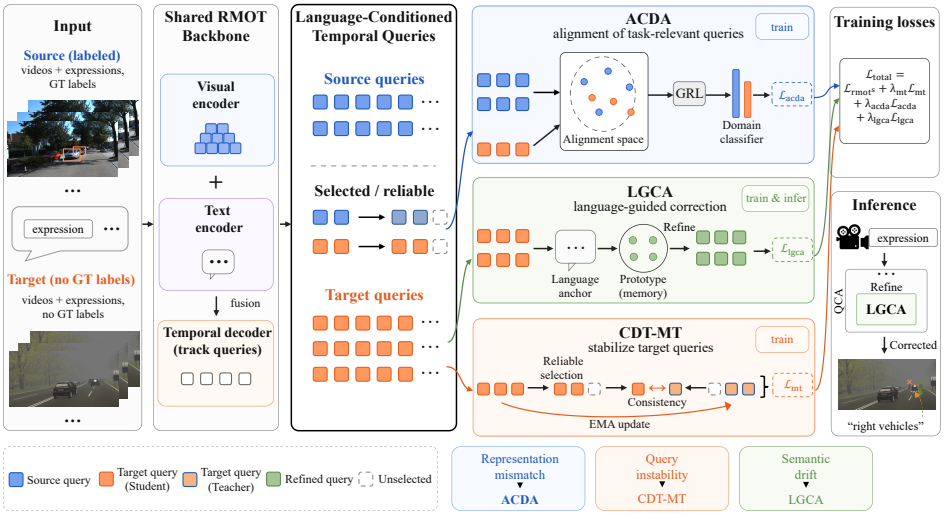


Figure 3: QCA component roles in UDA. CDT-MT supplies target-query consistency, ACDA aligns task-relevant queries, and LGCA provides language-guided calibration. Human tracking labels supervise the source branch; the diagram summarizes module motivations.

GRL is the identity in the forward pass and multiplies the feature gradient by $-\lambda_e$. The discriminator minimizes cross-entropy, while the student receives its reversed gradient. The feature-side coefficient is zero for the first five epochs and 0.1 thereafter. Domain entropy, source-prototype distance, and task confidence define the detached curriculum weights ω_i (Appendix B.2).

LGCA: language-guided calibration. Language-Guided Continuous Adaptation uses a pooled text anchor ℓ and query confidence $\kappa_i = \frac{1}{2}[1 + \cos(q_i, \ell)]u_i$. Let A contain queries with $\kappa_i > 0.7$, falling back to the top eight available queries when none qualify. Their weighted mean initializes a detached prototype p; subsequent update events apply

$$\widehat{p} = \frac{\sum_{i \in A} \kappa_i q_i}{\sum_{i \in A} \kappa_i + \varepsilon}, \qquad p \leftarrow 0.9p + 0.1\,\mathrm{sg}(\widehat{p}), \qquad x_i = q_i + 0.1 C_\xi(p - \overline{q}), \tag{4}$$

where $\overline{q}$ is the current query mean and C_ξ is a learned bottleneck MLP. The prototype provides a stream-level reference, while the offset correction preserves each query's individual feature through a residual. Normalized appearance and motion text anchors ℓ_a, ℓ_m condition two feature branches, combined as

$$q_i^* = g_a(x_i \odot \ell_a) + (1 - g_a) W_m(x_i \odot \ell_m). \tag{5}$$

Here $\odot$ denotes elementwise multiplication, W_m is a learned projection, and g_a is a detached, prototype-quality-dependent affine coefficient. Motion cues come from the expression; temporal history remains in the base tracker. During training, alignment to ℓ and consistency with $\mathrm{sg}(q_i)$ regularize source queries. Appendix B.3 defines the gate and auxiliary terms.

Inference and state. The student tracker runs with fixed network parameters on each video–expression stream. LGCA resets its prototype at the start of the stream and updates it from predicted query confidence every three frames. Its feature-level residual is

$$q_i^{\mathrm{out}} = (1 - \beta) q_i + \beta q_i^*, \qquad \beta = 0.15\,\mathrm{clip}\left(\frac{\zeta - 0.05}{0.45}, 0, 1\right), \tag{6}$$

where ζ is the stream's prototype–query quality and β is shared by its current queries. Non-finite quality disables the correction. The next-frame state is $S_t = U_\theta(\widetilde{S}_t, Q_t^{\mathrm{out}})$: query propagation U_θ consumes the tracker state $\widetilde{S}_t$ and calibrated queries. Current-frame boxes and identities remain the tracker readout. Output filtering uses objectness > 0.6, referring confidence > 0.4, and box area > 100 pixels. Appendix B.4 specifies the full transition.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We evaluate the six transfers defined in Sec. 3: clear to fog, rain, and 30°L; synthetic to Refer-KITTI-V2 and Refer-BDD; and Refer-KITTI-V2 to vKITTI fog. All settings use fixed splits and TrackEval.

Published RMOT baselines provide source-trained references; QCA and the query-alignment controls use UDA. The CoTTA-style comparator is a restricted query-state adaptation. MUTR and ReferDINO use the RVOS-to-RMOT interface in Appendix C.3.

We report TrackEval HOTA as the primary metric and DetA/AssA as diagnostic components. The three-seed TempRMOT/QCA runs use AdamW for 60 epochs on four RTX 3090 GPUs. Table 3 reports the single-run generic adaptations, with implementation details in Appendix C.2. QCA adds the objectives in Eq. (1) to source RMOT supervision.

5.2 MAIN RESULTS AND FINDINGS

The in-domain TempRMOT reference on *vKITTI clear → clear* reaches 28.69 HOTA, 17.38 DetA, and 47.39 AssA. The evaluated weather/viewpoint transfers have lower detection and association metrics (Fig. 1).

Table 2: Cross-domain HOTA/DetA/AssA for adapted RVOS, source-trained RMOT references, and QCA (UDA). The methods differ in target-data access. These are single-run results; Table 4 reports three-seed statistics.

Family	Method	vKITTI clear → fog			vKITTI clear → rain			vKITTI clear → 30°L		
		HOTA	DetA	AssA	HOTA	DetA	AssA	HOTA	DetA	AssA
RVOS	MUTR (Yan et al., 2024)	4.81	3.25	7.11	4.76	3.81	5.95	6.34	3.55	11.34
	ReferDINO (Liang et al., 2025)	5.88	4.19	8.25	5.58	4.41	7.07	6.78	4.03	11.43
RMOT	TransRMOT (Wu et al., 2023a)	11.67	4.10	33.95	11.91	3.95	36.67	9.99	3.06	33.53
	MGLT (Chen et al., 2025a)	10.37	4.09	26.36	12.40	4.49	34.39	9.73	2.47	38.58
	TempRMOT (Zhang et al., 2024)	16.00	8.79	29.20	19.50	10.30	37.05	20.43	10.08	41.79
	QCA (ours)	**21.37**	**11.04**	**41.38**	**24.20**	**13.48**	**43.49**	**25.25**	**14.43**	**44.31**

Family	Method	vKITTI clear → Refer-KITTI-V2			vKITTI clear → Refer-BDD			Refer-KITTI-V2 → vKITTI fog		
		HOTA	DetA	AssA	HOTA	DetA	AssA	HOTA	DetA	AssA
RVOS	MUTR (Yan et al., 2024)	5.33	2.38	11.90	4.91	2.45	9.86	3.48	2.15	5.63
	ReferDINO (Liang et al., 2025)	5.85	2.64	12.94	6.29	3.02	13.09	4.28	2.68	6.84
RMOT	TransRMOT (Wu et al., 2023a)	9.87	3.10	31.79	12.74	4.33	38.52	10.59	4.56	24.64
	MGLT (Chen et al., 2025a)	12.88	4.95	33.84	13.04	5.87	28.96	11.28	4.95	25.73
	TempRMOT (Zhang et al., 2024)	17.17	7.97	37.60	14.32	6.29	33.95	11.47	**5.44**	24.57
	QCA (ours)	**18.92**	**8.76**	**41.32**	**19.09**	**8.17**	**44.91**	**14.86**	5.10	**43.98**

Table 2 compares adapted RVOS and native RMOT methods across all six transfers. Appendix C.3 specifies the RVOS interface.

Among the controlled variants, *fog* yields the lowest TempRMOT HOTA, falling from the 28.69 clear→clear reference to 16.00 in the representative matrix. Both DetA and AssA decrease, showing that the shift affects the evaluated trajectories along detection- and association-sensitive dimensions.

QCA has the highest HOTA and AssA across the six evaluated transfers. On rain and viewpoint transfer, it improves HOTA over the source-trained TempRMOT reference by 4.70 and 4.82 points, respectively. These comparisons quantify the benefit of UDA over the no-adaptation reference; Table 3 compares adaptation mechanisms with target access.

5.3 Generic Adaptation and Reliability

Table 3: Generic adaptation controls on Clear→Fog. Single-run results on the same held-out evaluation set; regimes specify target-data access. Implementation scope is described in Appendix C.2.

Regime	Method	HOTA	DetA	AssA
UDA	Query-DANN (Ganin et al., 2016)	16.79	8.52	33.09
UDA	DANN+MeanTeacher (Deng et al., 2021)	17.04	8.74	33.21
Query TTA	CoTTA-style (Wang et al., 2022)	16.14	7.18	36.28
UDA	QCA	**21.37**	**11.04**	**41.38**

Query-DANN applies adversarial alignment to object queries; DANN+MeanTeacher adds an EMA teacher and confidence-filtered consistency. QCA exceeds these controls by 4.58 and 4.33 HOTA points in the evaluated configurations (Table 3). The CoTTA-style query-state control yields 36.28 AssA, 7.18 DetA, and 16.14 HOTA. The component metrics expose different detection–association tradeoffs; qualitative cases separately examine target selection.

Across three seeds, QCA improves HOTA over the TempRMOT reference by 4.82 points on fog and 4.72 on Refer-BDD on average, with gains in DetA and AssA (Table 4). All three seeds enter the summary statistics; Appendix C.1 specifies aggregation.

Table 4: Mean ± sample standard deviation across all three seeds (0, 1, 2), on held-out fog and Refer-BDD evaluation sets.

Target	Method	HOTA	DetA	AssA
Fog	TempRMOT	16.27 ± 0.89	8.75 ± 0.51	30.31 ± 1.65
	QCA	$\mathbf{21.09 \pm 0.25}$	$\mathbf{10.78 \pm 0.24}$	$\mathbf{41.27 \pm 0.10}$
Refer-BDD	TempRMOT	13.75 ± 0.50	5.60 ± 0.72	34.76 ± 3.54
	QCA	$\mathbf{18.48 \pm 0.55}$	$\mathbf{7.63 \pm 0.47}$	$\mathbf{44.90 \pm 0.29}$

QCA improves HOTA in every evaluated seed, with gains spanning 3.81–5.45 points on fog and 4.65–4.77 on Refer-BDD. Its AssA sample standard deviations are 0.10 and 0.29, compared with 1.65 and 3.54 for TempRMOT. These results show consistent improvements across the evaluated initializations in both controlled weather and synthetic-to-real transfer.

5.4 Ablation and Qualitative Analysis

Table 5: Component ablation on vKITTI clear→fog and clear→Refer-BDD. Each cell reports HOTA / DetA / AssA from the single-run ablation grid.

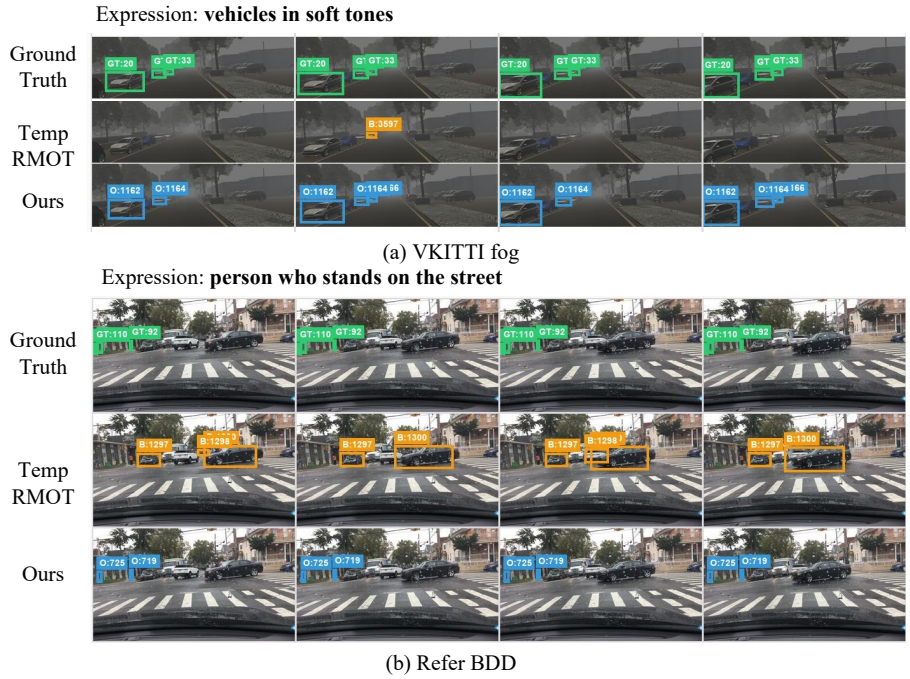
Components	Fog	Refer-BDD
TempRMOT	16.00 / 8.79 / 29.20	14.32 / 6.29 / 33.95
LGCA	16.17 / 8.62 / 30.34	15.14 / 6.21 / 36.80
ACDA	16.34 / 8.25 / 32.36	14.79 / 5.90 / 37.05
CDT-MT	16.25 / 8.48 / 31.14	14.61 / 6.03 / 35.35
CDT-MT + ACDA	16.58 / 7.92 / 34.89	15.69 / 6.34 / 38.83
CDT-MT + LGCA	17.59 / 8.19 / 38.02	16.73 / 6.92 / 40.50
ACDA + LGCA	18.73 / 8.98 / 39.08	17.84 / 7.46 / 42.70
Full QCA	**21.37 / 11.04 / 41.38**	**19.09 / 8.17 / 44.91**

Table 5 shows the strongest HOTA with all three components on both targets. The two-component variants reveal interactions among temporal consistency, domain alignment, and language calibration.

The strongest two-component variant, ACDA+LGCA, reaches 18.73 HOTA on fog and 17.84 on Refer-BDD. Adding CDT-MT increases these to 21.37 and 19.09, with DetA gains of 2.06 and 0.71 and AssA gains of 2.30 and 2.21. CDT-MT alone yields smaller HOTA improvements of 0.25 and 0.29 over the ablation baseline. Within this grid, temporal consistency is most effective when coupled

with task-relevant alignment and language calibration. LGCA-containing pairs also outperform CDT-MT+ACDA on both targets, supporting the role of expression-conditioned calibration in the combined system.

What changes with adaptation? Adaptation has domain-dependent precision–recall effects (Appendix C.4). On rain, DetPr rises from 16.64 to 41.93 while DetRe falls from 20.95 to 16.44. In reverse transfer, AssA rises from 24.57 to 43.98 while DetA falls from 5.44 to 5.10. The component metrics expose tradeoffs alongside the HOTA gains.



(a) VKITTI fog

(b) Refer BDD

Figure 4: Representative fog and Refer-BDD cases. The ground truth, TempRMOT, and QCA rows show how visually plausible predictions can drift from the referred targets under domain shift. Additional cases appear in Appendix C.5.

Expression-conditioned behavior. Fig. 4 shows two target-selection difficulties. In fog, "vehicles in soft tones" requires appearance-based selection among low-contrast vehicles. TempRMOT follows a different vehicle; QCA maintains the two referred targets across the displayed frames. On Refer-BDD, "person who stands on the street" combines category and spatial cues. TempRMOT outputs vehicle tracks, while QCA follows the annotated pedestrians. HOTA, DetA, and AssA score the resulting trajectories; these annotated examples examine expression consistency. Additional viewpoint and motion-dependent cases appear in Appendix C.5.

6 Conclusion

CD-RMOT-Bench evaluates unsupervised domain adaptation for RMOT across controlled weather/viewpoint variants and synthetic–real settings. Scene-local identity mapping, visibility-aware transfer, and candidate review make its digital-twin core auditable. QCA provides a query-based UDA reference, with gains over source-trained references across the evaluated transfers and initializations. Detection and association metrics expose domain-dependent tradeoffs, while qualitative cases illustrate target-selection errors. Together, component ablations and multi-seed comparisons complement the aggregate results by examining component combinations and variability across training initializations. The benchmark connects paired driving-scene analysis with cross-dataset evaluation; extension to broader scenes and capture conditions remains a direction for future work.

AI USE STATEMENT

Generative AI tools were used to polish the manuscript's language and improve its readability. All AI-assisted content was reviewed and verified by the authors. The authors take full responsibility for the final content of the paper.

REFERENCES

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

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

Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. In *CVPR*, 2022.

Dhanajit Brahma and Piyush Rai. A probabilistic framework for lifelong test-time adaptation. In *CVPR*, 2023.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual KITTI 2. *arXiv preprint arXiv:2001.10773*, 2020.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *CVPR*, 2019.

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

Tzoulio Chamiti, Leandro Di Bella, Adrian Munteanu, and Nikos Deligiannis. ReferGPT: Towards zero-shot referring multi-object tracking. In *CVPRW*, 2025.

Chenglizhao Chen, Shaofeng Liang, Runwei Guan, Xiaolou Sun, Haocheng Zhao, Haiyun Jiang, Tao Huang, Henghui Ding, and Qing-Long Han. AerialMind: Towards referring multi-object tracking in UAV scenarios. In *AAAI*, 2026.

Jiajun Chen, Jiacheng Lin, Guojin Zhong, You Yao, and Zhiyong Li. Multigranularity localization transformer with collaborative understanding for referring multiobject tracking. *IEEE Transactions on Instrumentation and Measurement*, 74:1–13, 2025a.

Meilin Chen, Weijie Chen, Shicai Yang, Jie Song, Xinchao Wang, Lei Zhang, Yunfeng Yan, Donglian Qi, Yueting Zhuang, Di Xie, et al. Learning domain adaptive object detection with probabilistic teacher. In *ICML*, 2022.

Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. Temporal attentive alignment for large-scale video domain adaptation. In *ICCV*, 2019.

Sijia Chen, Yanqiu Yu, En Yu, and Wenbing Tao. ReaMOT: A benchmark and framework for reasoning-based multi-object tracking. *arXiv preprint arXiv:2505.20381*, 2025b.

Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive Faster R-CNN for object detection in the wild. In *CVPR*, 2018.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *CVPR*, 2021.

Yunhao Du, Cheng Lei, Zhicheng Zhao, and Fei Su. iKUN: Speak to trackers without retraining. In *CVPR*, 2024.

Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*, 2016.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, 2012.

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.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, et al. WILDS: A benchmark of in-the-wild distribution shifts. In *ICML*, 2021.

Guangyao Li, Siping Zhuang, Yajun Jian, Yan Yan, and Hanzi Wang. Language decoupling with fine-grained knowledge guidance for referring multi-object tracking. In *ICCV*, 2025.

Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *CVPR*, 2022.

Tianming Liang, Kun-Yu Lin, Chaolei Tan, Jianguo Zhang, Wei-Shi Zheng, and Jian-Fang Hu. ReferDINO: Referring video object segmentation with visual grounding foundations. In *ICCV*, 2025.

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*, 25(11):18964–18977, 2024.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Conditional adversarial domain adaptation. In *NeurIPS*, 2018.

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

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

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *ICML*, 2022.

Xingchao Peng, Ben Usman, Neela Kaushik, Dequan Wang, Judy Hoffman, and Kate Saenko. VisDA: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, 2019.

Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016.

German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016.

Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, 2018.

Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, 2019.

Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. TransTrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020.

Tao Sun, Mattia Segu, Janis Postels, Yuxuan Wang, Luc Van Gool, Bernt Schiele, Federico Tombari, and Fisher Yu. SHIFT: A synthetic driving dataset for continuous multi-task domain adaptation. In *CVPR*, 2022.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021.

Likai Wang, Xiangqun Zhang, Ruize Han, Yanjie Wei, Song Wang, and Wei Feng. A new benchmark and algorithm for clothes-changing video person re-identification. *IEEE Transactions on Information Forensics and Security*, 20:1993–2005, 2025.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *CVPR*, 2022.

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

Dongming Wu, Wencheng Han, Tiancai Wang, Xingping Dong, Xiangyu Zhang, and Jianbing Shen. Referring multi-object tracking. In *CVPR*, 2023a.

Dongming Wu, Tiancai Wang, Yuang Zhang, Xiangyu Zhang, and Jianbing Shen. OnlineRefer: A simple online baseline for referring video object segmentation. In *ICCV*, 2023b.

Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *CVPR*, 2022.

Bin Yan, Yi Jiang, Peize Sun, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Towards grand unification of object tracking. In *ECCV*, 2022.

Shilin Yan, Renrui Zhang, Ziyu Guo, Wenchao Chen, Wei Zhang, Hongyang Li, Yu Qiao, Hao Dong, Zhongjiang He, and Peng Gao. Referred by multi-modality: A unified temporal transformer for video object segmentation. In *AAAI*, 2024.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020.

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

Xiangqun Zhang, Ruize Han, Likai Wang, Linqi Song, Junhui Hou, and Wei Feng. Synthetic-to-real video person Re-ID. *IEEE Transactions on Information Forensics and Security*, 20:5438–5450, 2025a.

Xiao Zhang, Fei Wei, Yong Wang, Wenda Zhao, Feiyi Li, and Xiangxiang Chu. UPRE: Zero-shot domain adaptation for object detection via unified prompt and representation enhancement. In *ICCV*, 2025b.

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

Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. FairMOT: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129:3069–3087, 2021.

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

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

Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *ECCV*, 2020.

A Benchmark Construction and Protocols

A.1 Alignment, Matching, and Residual Calibration

Refer-vKITTI is constructed from the five Refer-KITTI-V2 sequences that have aligned Virtual KITTI 2 counterparts. For each paired real/synthetic scene, the construction pipeline loads the Refer-KITTI-V2 frame-level labels and the vKITTI bounding boxes, converts them into a common box representation, and enumerates source–target identity candidates after scene and frame alignment. This formulation turns expression transfer into a constrained identity-mapping problem rather than an unconstrained annotation task, which is essential for preserving referential semantics under controlled domain changes.

Candidate evidence is collected through frame-level matching with an IoU threshold of 0.02. This threshold is intentionally permissive because it is used to collect correspondence evidence before scene-level assignment, not as a standalone acceptance rule. For each source–target identity pair, the pipeline computes frame-match count, mean frame-level IoU, direct track IoU over overlapping frames, center similarity, area similarity, and temporal IoU. These terms define the per-scene matching score

$$s(i,j) = 4N_{ij} + 30\bar{I}_{ij}^{\text{frame}} + 6\bar{I}_{ij}^{\text{track}} + 2C_{ij} + A_{ij} + 3T_{ij}, \tag{7}$$

where N_{ij} is the number of IoU-filtered frame matches, $\bar{I}_{ij}^{\text{frame}}$ is the mean frame-level IoU, $\bar{I}_{ij}^{\text{track}}$ is the direct track IoU over overlapping frames, C_{ij} and A_{ij} denote center and area similarity, and T_{ij} denotes temporal IoU. The resulting scene-level identity correspondences are selected under a one-to-one assignment constraint, yielding auditable per-scene KITTI-to-vKITTI identity mappings.

The resulting mappings are then used to transfer Refer-KITTI-V2 expression-track annotations to all vKITTI domains. During normalization, the natural-language expression is preserved, video identifiers are rewritten to the corresponding vKITTI domain, target identities are remapped according to the per-scene correspondence, and frame-level targets are removed when the mapped vKITTI identity is absent in the target-domain frame. This accounts for visibility changes, viewpoint-induced invalidation, and domain-specific object absence, while keeping the final annotations compatible with the unified RMOT evaluation protocol.

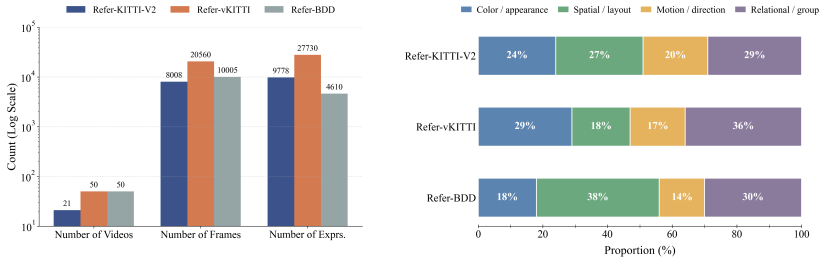
The candidate audit covers all 374 IoU-supported pairs: two independent reviewers initially agreed on 369, and five disagreements were adjudicated using complete trajectories. The final 251/123 accepted/rejected partition was unchanged. A separate stratified audit checked 300 transferred expression–track instances, including 100 double-reviewed cases (96% initial agreement). These figures measure inter-annotator agreement and review coverage. Table 6 reports the construction counts using their respective units.

Table 6: Refer-vKITTI construction statistics. Expression–track instances, identity-pair hypotheses, and frame–target mentions are distinct counting units. The complete semantic audit is reported in Table 7.

Item	Value	Evidence / Notes
Total transferred videos	50	Five aligned KITTI/vKITTI scenes expanded to 10 vKITTI domains
Total transferred expressions	27,730	Final Refer-vKITTI expression directory
Total transferred expression–track pairs	285,160	Source expression–track pairs expanded across 10 target domains
Candidate identity pairs after alignment	6,887	Per-scene source–target identity pairs with temporal frame overlap
Candidate pairs with IoU evidence	374	Frame-level candidate evidence collected with IoU ≥ 0.02
Matched identity pairs after one-to-one assignment	251	Per-scene identity mappings summed over five aligned scenes
Unmatched source identities in build summary	0	Reported by scene-level mapping summaries
Frame-level visibility invalidations	806,110	Dropped target mentions when mapped vKITTI identities are absent in target-domain frames
Discarded expression–track pairs	13,117	Initial transferred pairs minus final retained pairs
Final retained expression–track pairs	272,043	Expression–track pairs retained after visibility filtering and normalization
Final retention rate	95.40%	Final retained pairs divided by initially transferred pairs
IoU threshold for evidence collection	0.02	Used before scene-level assignment, not as a standalone acceptance rule
Per-scene matching score	Composite score	Frame-match count, frame IoU, direct track IoU, center similarity, area similarity, and temporal IoU
Temporal consistency cue	Soft score term	Temporal IoU contributes to the matching score
Lightweight consistency check	306,505 checks, 0 flagged	Post-construction consistency validation on generated labels

Table 7: Construction audit: candidate coverage, inter-annotator agreement, and transfer review.

Identity-pair candidate audit	Expression–track transfer audit
All 374/374 supported candidates independently reviewed by two auditors; 369 initial agreements (98.66%, $\kappa = 0.970$); five disagreements adjudicated from complete trajectories; frozen partition 251 accepted / 123 rejected, unchanged after audit.	300 stratified transferred instances sampled across weather/viewpoint and expression cues; 100 double-reviewed with 96% initial agreement; discrepancies resolved with temporal/context review.



(a) Dataset Statistics (b) Approximate Attribute Composition

Figure 5: Benchmark composition and expression-cue characterization using automatically counted, non-exclusive keyword categories.

A.2 Splits, Protocols, and Implementation

For controlled vKITTI transfer, we use three clear-domain Refer-vKITTI scenes for source training, containing 1,407 frames and 1,710 expressions. Each reported target setting, including *fog*, *rain*, and *30°L*, uses two held-out aligned scenes with 659 frames and 1,063 expressions. The real clear-domain setting follows the same cross-scene principle in Refer-KITTI-V2, using three videos for training with 1,359 frames and 1,710 expressions and two held-out videos for evaluation with 646 frames and 1,063 expressions. For synthetic-to-real transfer, Refer-BDD provides 8,385 training frames and 3,937 expressions for adaptation, and 1,619 test frames with 673 expressions for evaluation.

Refer-vKITTI uses aligned digital-twin counterparts within the source/target adaptation scenes, enabling controlled weather and viewpoint comparisons. The held-out evaluation scenes are disjoint from these training-side scenes: source clear and target adaptation use Scene06/18/20, while fog evaluation uses Scene01/02. The design is therefore scene-disjoint at evaluation, although source and

target adaptation intentionally share underlying scene layouts. We complement it with synthetic-to-real and real-to-synthetic experiments involving Refer-KITTI-V2 and Refer-BDD.

Table 8: Access boundary for scene-disjoint inductive UDA. "Eval only" means held by TrackEval, not visible to model training or checkpoint selection.

Information	Source train	Target adapt	Target eval
Video and referring sentence	Yes	Yes	Inference input
Boxes, track IDs, expression–object links	Yes	No	Eval only
Real–virtual mapping / paired correspondence	Construction only	No	No
Supervised RMOT loss	Yes	No	No
Teacher predictions / query consistency	No	Yes	No

Table 8 specifies the UDA inputs. The TempRMOT reference uses no target adaptation data; the query-TTA comparator updates stream-local state during inference (Appendix C.2).

For the three-seed TempRMOT/QCA experiments we use AdamW for 60 epochs, with learning-rate decay after epoch 40 and batch size one per GPU on four NVIDIA RTX 3090 GPUs. QCA uses EMA decay 0.999, ACDA top-$M = 16$ queries, and LGCA prototype momentum 0.9. The generic adaptations use the separate single-run configurations described in Appendix C.2.

B Reference Method Details

We separate the tracker parameters θ, the domain discriminator parameters ϕ, and the LGCA correction parameters ξ. The EMA teacher $\bar{\theta}$ follows the student tracker. In the following, sg stops gradients, $\varepsilon = 10^{-6}$ stabilizes normalization, and $e \in \{0, \dots, E_{\max} - 1\}$ is the zero-based epoch. Objectness and referring scores are sigmoid probabilities. All box consistency terms use normalized (c_x, c_y, w, h) coordinates.

B.1 CDT-MT: Reliable Query Selection

Teacher pseudo targets and query indexing. The teacher processes the target clip without gradient updates. Per frame, its objectness–referring confidence product ranks pseudo boxes, which supply the student's target-side frame matching and query bookkeeping. The supervised RMOT criterion is evaluated on source annotations. The target auxiliary loss is the consistency objective in Eq. 2. It operates on the final-frame exported query arrays, using their shared index range $N = \min(N_{\mathrm{stu}}, N_{\mathrm{tea}})$. This slot-wise consistency is separate from the tracker's frame-level Hungarian matching against teacher pseudo boxes.

Reliability and agreement. For each shared slot i, teacher reliability is $\rho_i = u_i^{\mathrm{tea}} v_i^{\mathrm{tea}}$. Detached student–teacher predictions define

$$a_i^{\mathrm{feat}} = \tfrac{1}{2}\big[1 + \cos(q_i^{\mathrm{stu}}, q_i^{\mathrm{tea}})\big], \tag{8}$$

$$a_i^{\mathrm{box}} = \exp\big[-4\,\mathrm{MAE}(b_i^{\mathrm{stu}}, b_i^{\mathrm{tea}})\big]\,, \tag{9}$$

$$a_i^{\mathrm{ref}} = 1 - |v_i^{\mathrm{stu}} - v_i^{\mathrm{tea}}|, \qquad a_i = \tfrac{1}{3}(a_i^{\mathrm{feat}} + a_i^{\mathrm{box}} + a_i^{\mathrm{ref}}), \tag{10}$$

where $\mathrm{MAE}(b, b') = \frac{1}{4}\sum_{k=1}^{4} |b_k - b'_k|$. Thus the factor 4 multiplies a coordinate mean, not an unnormalized L_1 norm. The ranking score is $\psi_i = 0.65\rho_i + 0.35a_i$.

Selection and weights. Let $C = \{i : \rho_i \geq 0.6\}$ and $k = \min\{N, 64, \max[8, \mathrm{round}(0.1N)]\}$. If $|C| \geq 8$, select the highest-scoring $\min(k, |C|)$ slots in C; otherwise, select the top k from all N slots. Pseudo-box construction uses the same count and reliability threshold with ρ_i as its ranking score. The consistency subset uses ψ_i and weights

$$w_i = \frac{\mathrm{sg}(\psi_i)\mathbf{1}[i \in \Omega_t]}{\sum_{j \in \Omega_t} \mathrm{sg}(\psi_j) + \varepsilon}. \tag{11}$$

An empty query set contributes zero loss. The feature MSE averages over d coordinates; smooth-L_1 averages over four box coordinates and uses unit transition. Their coefficients are 1 and 0.5, with coefficient 0.5 for squared referring-probability difference. Selection and weight computation receive no gradient. The warmup factor $\eta_e = \min(1, (e+1)/5)$ scales the complete target consistency loss.

Update order. Each training iteration runs the source student, target teacher, and target student branches, computes their auxiliary objectives, and takes one student/discriminator/calibrator optimizer step. The teacher then updates as $\bar{\theta} \leftarrow 0.999\bar{\theta} + 0.001\theta$. Teacher predictions are fixed targets within that iteration; no target human annotations enter this update.

B.2 ACDA: Adversarial Alignment and Curriculum

Each source or target clip contributes up to 16 final-frame queries ranked by $c_i = u_i v_i$. The selected features z_i retain their student gradients. The domain label d_i is a dataset-domain identifier, distinct from an object class or trajectory identity. The domain head is a $d \to 128 \to K_d$ MLP with ReLU and dropout 0.3; K_d is the configured number of domain logits. The default configuration has $K_d = 4$, even when a run uses only a subset of these domain labels.

For $\pi_i = \mathrm{softmax}(g_\phi(z_i))$, the difficulty is computed from detached values:

$$h_i^{\mathrm{ent}} = -\frac{\sum_{k=1}^{K_d} \pi_{ik} \log(\pi_{ik} + \varepsilon)}{\log K_d}, \tag{12}$$

$$h_i^{\mathrm{proto}} = \mathrm{norm}_{\mathcal{Q}}(\|z_i - \mu_s\|_2), \qquad h_i^{\mathrm{task}} = 1 - \mathrm{norm}_{\mathcal{Q}}(c_i), \tag{13}$$

$$h_i = \mathrm{clip}(0.5h_i^{\mathrm{ent}} + 0.25h_i^{\mathrm{proto}} + 0.25h_i^{\mathrm{task}}, 0, 1). \tag{14}$$

Here $\mathrm{norm}_{\mathcal{Q}}(x_i) = (x_i - \min_j x_j)/(\max_j x_j - \min_j x_j + \varepsilon)$, with value zero for a constant set. The detached source prototype μ_s is initialized from the source-query mean and subsequently updated with momentum 0.9. The curriculum weights are

$$\omega_i = 0.2 + 0.8\big[p_e + (1 - p_e)(1 - h_i)\big], \qquad p_e = \frac{e}{E_{\max} - 1}. \tag{15}$$

The weights gradually approach one, so difficult queries are downweighted most strongly early in training. They are averaged with cross-entropy over $|\mathcal{Q}|$ as in Eq. 3, without a second normalization by $\sum_i \omega_i$.

The gradient reversal satisfies $\mathrm{GRL}_{\lambda_e}(z) = z$ and $\partial\,\mathrm{GRL}\,/\partial z = -\lambda_e I$, with $\lambda_e = 0$ for $e < 5$ and 0.1 thereafter. The domain classifier still learns during this feature-gradient warmup. The outer loss coefficient 0.1 in Eq. 1 and the GRL coefficient have different roles. In the memory-saving ACDA-without-CDT-MT ablation, target features are detached reference features; source features and the discriminator remain trainable.

B.3 LGCA: Prototype, Fusion, and Training Objective

Language-conditioned prototype. Let ℓ be the normalized mean token embedding. For objectness u_i, set $\kappa_i = \frac{1}{2}[1 + \cos(q_i, \ell)]u_i$. Select $A = \{i : \kappa_i > 0.7\}$; if empty, use the top $\min(8, N)$ queries. At each update event,

$$\widehat{p} = \frac{\sum_{i \in A} \kappa_i q_i}{\sum_{i \in A} \kappa_i + \varepsilon}, \qquad p \leftarrow \begin{cases} \mathrm{sg}(\widehat{p}), & \text{first update,} \\ 0.9p + 0.1\,\mathrm{sg}(\widehat{p}), & \text{otherwise.} \end{cases} \tag{16}$$

Prototype accumulation is detached. During training it is a running source-query state; during inference it is reset for each video–expression pair and updated every three frames. For an initialized prototype, the correction is

$$x_i = q_i + 0.1C_\xi(p - \overline{q}), \qquad \overline{q} = \frac{1}{N}\sum_{j=1}^{N} q_j, \tag{17}$$

where C_ξ is a $d \to d/4 \to d$ MLP with ReLU. Before initialization, $x_i = q_i$.

Appearance and motion fusion. Pooling appearance-related and motion-related tokens yields normalized anchors ℓ_a and ℓ_m; unavailable token groups use the whole-expression mean. The feature-level branches are $q_i^{\text{app}} = x_i \odot \ell_a$ and $q_i^{\text{mot}} = W_m(x_i \odot \ell_m)$. Motion here denotes language cues; temporal history is maintained by the base tracker. The prototype quality is $\zeta = \frac{1}{N}\sum_i \cos(x_i, p)$, with initial value 0.5. The quality-conditioned affine gate is

$$g_a(\zeta) = \begin{cases} \text{sg}(\zeta), & \zeta < 0.5, \\ 0.8, & \zeta \geq 0.5, \end{cases} \qquad g_m = 1 - g_a, \qquad q_i^* = g_a q_i^{\text{app}} + g_m q_i^{\text{mot}}. \tag{18}$$

The gate coefficients sum to one; the raw cosine quality may be negative, so the expression is an affine combination. The outer inference coefficient in Eq. 6 is clipped to $[0, 0.15]$ and disables the correction when $\zeta \leq 0.05$.

Source-query auxiliary loss. The training objective is

$$\begin{aligned} \mathcal{L}^s_{\text{LGCA}} = 1 - \frac{1}{N}\sum_i \cos(q_i^*, \ell) + \frac{0.5}{Nd}\sum_i \|q_i^* - \text{sg}(q_i)\|_2^2 \\ - 0.1\zeta + 0.1\|(g_a, g_m) - (0.5, 0.5)\|_1. \end{aligned} \tag{19}$$

The text-alignment and quality terms update the query representation and correction parameters. The consistency target is detached. Gate coefficients are detached, making the final gate term a scalar diagnostic within the objective rather than a gradient-trained gate penalty. The auxiliary term is applied to the source clip with outer weight 0.1.

B.4 Inference State and Output

The inference formulation uses the student tracker with fixed parameters. Let S_{t-1} contain the incoming track-query state, identities, reference points, and history. Denote its frame forward/readout by F_θ and query propagation by U_θ. For frame I_t and expression E, the operation order is

$$\begin{aligned} (B_t, R_t, \mathbf{u}_t, Q_t, \widetilde{S}_t) &= F_\theta(I_t, E, S_{t-1}), \\ (Q_t^{\text{out}}, p_t) &= \text{LGCA}_\xi(Q_t, E, \mathbf{u}_t, p_{t-1}; t), \\ S_t &= U_\theta(\widetilde{S}_t, Q_t^{\text{out}}). \end{aligned} \tag{20}$$

B_t, R_t, and $\mathbf{u}_t$ denote current boxes, referring scores, and objectness probabilities; Q_t contains query features. The residual supplies the next-frame query-state update, after the current prediction readout. Current boxes retain the tracker predictions. On frames without a prototype update, $p_t = p_{t-1}$; an empty query set leaves the prototype unchanged.

At the first frame of each video–expression pair, track and prototype state are reset. Prototype updates use predicted confidence; no optimizer step or target label is required. Teacher and discriminator predictions are unnecessary at inference. Non-finite or low prototype quality leaves queries unchanged. The base tracking procedure maintains identities, and the final output keeps tracks with valid identities, objectness above 0.6, referring score above 0.4, and area above 100 image pixels.

Equations 1 and 20 specify UDA training and fixed-parameter tracking with stream-local state updates, respectively. The reported comparisons evaluate the complete QCA system.

C Additional Experimental Results

C.1 Multi-Seed Aggregation and Run Definitions

Table 4 reports the arithmetic mean and sample standard deviation over seeds 0, 1, and 2 for each method and target. Every seed is included. We present aggregate statistics rather than a separate per-seed table. The HOTA improvements over TempRMOT average 4.82 points on fog and 4.72 on Refer-BDD.

The single-run comparison and component-ablation grid use the fog TempRMOT baseline of 16.00 HOTA. The multi-seed evaluation is a separate run set. Extended detection and association metrics in Table 9 belong to the comparison grid. Each result is interpreted with its own run configuration.

C.2 Generic Adaptation Implementations

Table 3 reports single-run query-level adaptations of generic mechanisms on Clear→Fog. Query-DANN uses a binary discriminator to align source and target query features. DANN+MeanTeacher adds an EMA teacher and confidence-filtered pseudo-target consistency. QCA additionally uses query ranking, curricular alignment, and language-guided calibration.

The CoTTA-style control updates a per-expression query prototype online, with restoration toward a stream anchor and fixed network weights. This query-state adaptation implements a restricted mechanism inspired by CoTTA; the original algorithm also adapts network weights. The controls use separately configured training budgets, so the table describes the evaluated configurations rather than an equal-compute comparison. Their single-run scope complements the three-seed TempRMOT/QCA evaluation.

C.3 Adapted RVOS Baselines

RVOS methods predict language-conditioned mask sequences rather than RMOT trajectories. To include them as cross-task references, we adapt MUTR and ReferDINO through a pseudo-mask interface. For each video-expression pair, the RMOT bounding boxes associated with the referred track identities are rasterized into rectangular pseudo masks, yielding RVOS-style training samples without introducing extra mask annotation. The RVOS models are trained under the same source-domain splits and applied to the corresponding target-domain videos.

The available inference adapters select the highest-scoring query within each temporal window, threshold its masks, and convert each mask to its tight enclosing box. Empty masks produce missed detections; disconnected foreground regions share one enclosing box. The adapters export per-expression trajectories in MOT format. This interface evaluates language-conditioned mask transfer with single-query selection and does not provide independent identities for multiple objects within one selected mask.

C.4 Extended Metric Decompositions

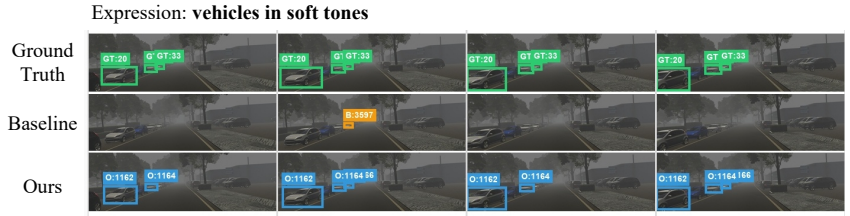
Table 9: Extended metrics for the single-run comparison grid. Parentheses report changes relative to the TempRMOT row in the same setting. Run definitions are given in Appendix C.1.

Setting	Method	HOTA	Detection			Association			LocA
			DetA	DetRe	DetPr	AssA	AssRe	AssPr	
vKITTI clear → fog	TempRMOT	16.00	8.79	14.20	18.58	29.20	34.45	70.61	**88.87**
	QCA	21.37 (+5.37)	11.04 (+2.25)	18.02 (+3.82)	21.80 (+3.22)	41.38 (+12.18)	47.86 (+13.41)	72.10 (+1.49)	87.14 (-1.73)
vKITTI clear → rain	TempRMOT	19.50	10.30	20.95	16.64	37.05	43.49	68.89	86.39
	QCA	24.20 (+4.70)	13.48 (+3.18)	16.44 (-4.51)	41.93 (+25.29)	43.49 (+6.44)	51.04 (+7.55)	71.11 (+2.22)	89.90 (+3.51)
vKITTI clear → 30°L	TempRMOT	20.43	10.08	23.92	14.65	41.79	53.74	67.46	85.45
	QCA	25.25 (+4.82)	14.43 (+4.35)	19.85 (-4.07)	33.77 (+19.12)	44.31 (+2.52)	49.76 (-3.98)	78.42 (+10.96)	88.21 (+2.76)
vKITTI clear → Refer-KITTI-V2	TempRMOT	17.17	7.97	21.33	11.03	37.60	47.52	61.29	78.76
	QCA	18.92 (+1.75)	8.76 (+0.79)	15.63 (-5.70)	16.11 (+5.08)	41.32 (+3.72)	49.19 (+1.67)	67.68 (+6.39)	79.28 (+0.52)
vKITTI clear → Refer-BDD	TempRMOT	14.32	6.29	10.88	12.61	33.95	41.19	62.45	77.16
	QCA	19.09 (+4.77)	8.17 (+1.88)	9.13 (-1.75)	42.83 (+30.22)	44.91 (+10.96)	51.84 (+10.65)	72.68 (+10.23)	88.61 (+11.45)
Refer-KITTI-V2 → vKITTI fog	TempRMOT	11.47	5.44	8.46	13.00	24.57	28.98	62.66	82.40
	QCA	14.86 (+3.39)	5.10 (-0.34)	16.18 (+7.72)	6.83 (-6.17)	43.98 (+19.41)	54.74 (+25.76)	63.00 (+0.34)	78.45 (-3.95)

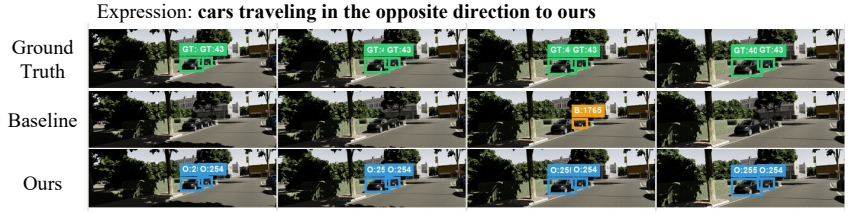
Table 9 decomposes the single-run comparison into detection, association, and localization metrics. The different precision–recall responses across domains complement the HOTA comparison in the main text.

C.5 Additional Qualitative Cases

Fig. 6 provides qualitative comparisons under weather, viewpoint, real-domain, and motion-dependent settings. In these examples, visually plausible detections can include objects outside the referred target set. The cases complement aggregate tracking metrics by showing which objects are selected across the displayed frames.



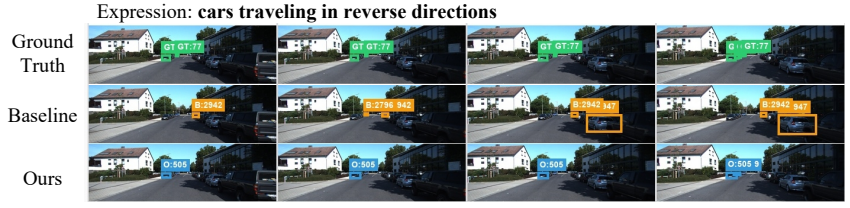
(a) Fog



(b) 30°L



(c) Refer-BDD



(d) Refer-KITTI-V2

Figure 6: Additional cross-domain RMOT cases. The displayed trajectories illustrate distractor selection and differences in expression-consistent tracking across weather, viewpoint, and real-domain settings.