# RMOT-COMPASS: A QUERY-CONDITIONED ERROR ATTRIBUTION FRAMEWORK FOR REFERRING MULTI-OBJECT TRACKING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Referring multi-object tracking (RMOT) tracks the objects described by a language query and preserves their identities over time. The query decides which scene objects are targets and which are distractors. Thus, a trajectory can be accurate and consistent, yet still wrong for the query. However, an aggregate score such as HOTA does not reveal how different output errors contribute to the recoverable gap. We argue that error diagnosis in RMOT should answer three distinct questions: how often an error occurs, what is gained by repairing it alone, and how much of the jointly recoverable gap it accounts for. In this work, we propose **RMOT-Compass**, a query-conditioned error attribution framework. It combines the language query with full-scene annotations to define target and distractor roles, and freezes diagnostic correspondences to keep error labels fixed across repairs. Joint repairs address five error types, covering missing targets, distractors, background or duplicate outputs, association, and localization, while coordinating target recovery with identity and box correction across frames. Standard Shapley allocation accounts for repair interactions by averaging marginal gains over all repair orders. The credited gains sum exactly to the jointly recoverable HOTA gap. Across five RMOT methods evaluated on the Refer-KITTI dataset, missing targets are consistently the largest recoverable error. Methods with higher overall scores can still have larger recoverable contributions from specific error types. Both findings hold on the Refer-KITTI-V2 dataset. Therefore, RMOT-Compass turns a single aggregate score into a query-aware error budget. It identifies which output errors account for the recoverable gap and directs focused model investigation.

## 1 INTRODUCTION

Referring multi-object tracking (RMOT) requires tracking every object described by a language query while preserving its identity across frames (Wu et al., 2023). Unlike conventional multi-object tracking, the query decides which scene objects count as targets. Fig. 1(a) shows the same car as a target under "vehicles on the left" and a distractor under "vehicles on the right". Therefore, a trajectory can be accurate and consistent, yet still wrong for the query.

Evaluation of the RMOT task adopts HOTA (Luiten et al., 2021), a standard metric for multi-object tracking. HOTA combines detection, association, and localization quality into a single score. A single score, however, does not say which output failures are behind it. In this work, we evaluate how correcting different combinations of output errors changes HOTA. The gain over the raw score from applying all repairs is the jointly recoverable gap. We argue that diagnosing an error requires three different questions. **How often does the error occur? How much score is gained by repairing this error alone, which we call the isolated gain? How much of the recoverable gap should be credited to this error when all repairs are applied together, which we call the joint attribution?** These questions have different answers. Error rates are computed over different reference populations. Fig. 1(b) shows that repair gains also depend on context: removing distractors from TransRMOT outputs gains 4.139 HOTA points alone but 6.716 after missing targets are restored. The five isolated gains sum to 40.645 points, below the 54.524 recovered jointly. A joint attribution is needed to account for this non-additive recovery.



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.

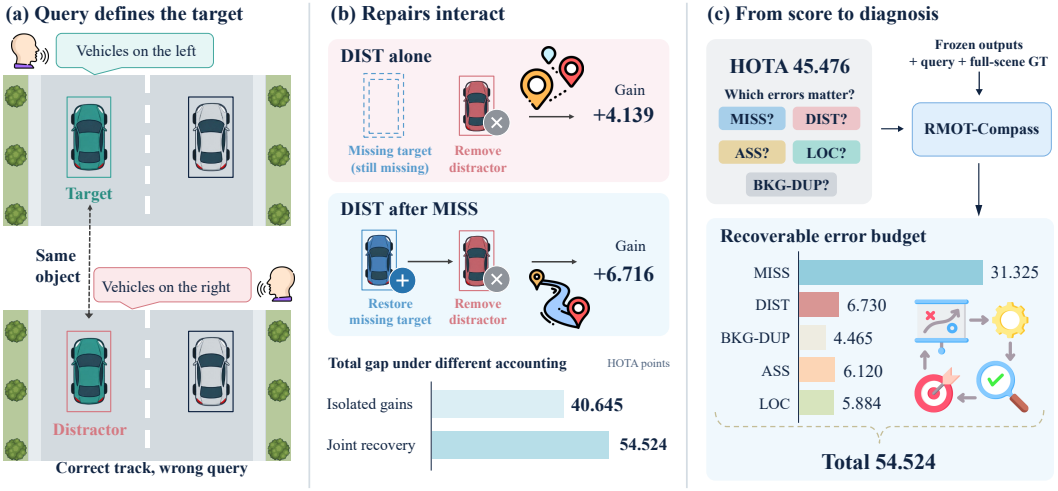



Figure 1: **Why RMOT needs query-conditioned error attribution.** (a) Changing the query changes the same object's role from target to distractor. (b) On TransRMOT, removing distractors gains 4.139 HOTA points alone and 6.716 after missing-target repair. Isolated gains sum to 40.645 points, below the 54.524 recovered jointly. (c) Given frozen outputs, the query, and full-scene ground truth, RMOT-Compass decomposes the 54.524-point recoverable HOTA gap above the raw score of 45.476 into five error types.

Diagnosing errors from outputs alone, without access to model internals, has precedent in object detection. TIDE (Bolya et al., 2020) defines detection error types and reports the gain from fixing each type in isolation. RMOT raises two problems that TIDE does not face. First, whether a valid scene object belongs in the output depends on the query, so the roles of target and distractor must be determined before any error can be counted. Second, restoring a missing target requires coordinating its identity with retained trajectories and with any association repair. The inserted target's identity depends on whether association repair is also selected. Any allocation rule can only divide scores that are already computed. It cannot determine these roles or build these repairs. The main challenge is to define query-conditioned repairs.

To address these questions, we propose **RMOT-Compass**, a query-conditioned error attribution framework. As illustrated in Fig. 1(c), it converts a method's recoverable HOTA gap into an error budget by quantifying output-error contributions under joint repair. Using the query and full-scene annotations, it distinguishes missing targets from tracked distractors and freezes correspondences between predictions and scene objects. Every repair remains tied to these original error labels. It defines five joint repairs for missing targets, distractors, background or duplicate outputs, identity, and localization. Every subset of these repairs produces a repaired output from the same raw predictions, and each output is scored with HOTA. Because repairs interact, each repair is credited with its average marginal gain over all orders in which the repairs can be applied. This credit is the Shapley value from cooperative game theory (Shapley et al., 1953). The credits sum exactly to the recoverable gap.

We apply RMOT-Compass to five RMOT methods on the Refer-KITTI dataset using released checkpoints without retraining. Missing targets receive the largest credit in every method, from 24.837 to 31.325 HOTA points. A higher HOTA score does not imply smaller credits for the other error types. Localization makes this concrete. Nearly every tracked target has some box error, so the localization error rate is 1.0 for every method, yet its credited HOTA impact ranges from 5.884 to 14.493 points. Furthermore, we evaluate RMOT-Compass on the Refer-KITTI-V2 dataset, demonstrating that the same error definitions, joint repair rules, and attribution procedure can be reused across datasets.

In summary, our main contributions are as follows:

- We propose RMOT-Compass, a query-conditioned error attribution framework for RMOT. It combines query-defined target and distractor roles with frozen diagnostic correspondences to identify five observable error types while keeping metric evaluation independent.

- We formulate temporally coordinated joint repairs that account for dependencies among target recovery, identity consistency, and localization. The framework distinguishes error occurrence, isolated gain, and joint attribution, using standard Shapley allocation to distribute the jointly recoverable HOTA gap while accounting for repair interactions.
- We conduct a comparative diagnosis of five RMOT methods on Refer-KITTI, complemented by controlled input perturbations and protocol reuse on Refer-KITTI-V2. This analysis identifies missing targets as the dominant recoverable component and exposes method-specific error contributions hidden by aggregate scores.

## 2 Related Work

**Referring multi-object tracking.** Referring multi-object tracking (RMOT) combines language-based object selection with temporal identity prediction. End-to-end methods integrate language guidance into tracking. TransRMOT introduces a transformer formulation and the Refer-KITTI dataset (Wu et al., 2023). TempRMOT enhances temporal object queries and introduces the Refer-KITTI-V2 dataset (Zhang et al., 2024). DKGTrack decouples appearance and motion (Li et al., 2025), HFF-Tracker uses hierarchical fusion (Zhao et al., 2025), CDRMT disentangles semantic queries (Liang et al., 2025), and COAL adds counterfactual supervision (Jia et al., 2026). VMRMOT aligns motion descriptions with vision and language (Lv et al., 2025), while STORM jointly performs grounding and tracking within a multimodal language model (Lu et al., 2026). Two-stage methods generate candidate trajectories and then select those matching the language query. iKUN unifies visual and textual knowledge (Du et al., 2024), FlexHook learns language-conditioned correspondence (Li et al., 2026), and ReferGPT uses spatial captions for zero-shot selection (Chamiti et al., 2025). Recent extensions broaden the task to cross-view (Chen et al., 2025), aerial (Chen et al., 2026), omnidirectional (Zhou et al., 2026), RGB–thermal (Yu et al., 2026), and cross-domain settings (Zhang et al., 2026).

**Tracking evaluation and error diagnosis.** The CLEAR MOT Metrics summarize detection and identity-switch errors (Bernardin & Stiefelhagen, 2008). IDF1 measures identity consistency through trajectory matching (Ristani et al., 2016), while HOTA balances detection and association across localization thresholds (Luiten et al., 2021). CAST adds temporal context to switch and transfer evaluation (Bai & Hager, 2026). In object detection and instance segmentation, TIDE quantifies the gain from correcting each output error type in isolation (Bolya et al., 2020). These approaches support performance measurement and output-error interpretation across tasks.

**Cooperative attribution and repair interactions.** Shapley values average marginal contributions over player orders (Shapley et al., 1953), and interaction indices summarize joint effects (Grabisch & Roubens, 1999). SHAP applies additive attribution to model predictions (Lundberg & Lee, 2017), while Sundararajan & Najmi (2020) show that different operational definitions of the game can produce different explanations. A cooperative attribution problem specifies its players and coalition values before assigning credit. Pairwise interaction indices further describe whether two players contribute more or less together than separately, averaged over the remaining coalition contexts.

RMOT-Compass extends output-level error diagnosis to errors whose meaning depends on a language query. Tracking metrics score trajectories against the selected targets but do not allocate the remaining gap to query-defined failures. TIDE measures isolated corrections in detection and segmentation, without the temporal identity constraints of RMOT. We define target and distractor roles, freeze their diagnostic correspondences, and construct joint repairs that coordinate target recovery with identity and box correction across frames. These RMOT-specific definitions determine the repair game. Standard Shapley allocation then distributes its recoverable gap among error types.

## 3 RMOT-Compass

### 3.1 Overview

We propose **RMOT-Compass**, a query-conditioned error attribution framework, as shown in Fig. 2. It takes frozen predictions, a language query, and full-scene ground truth. It assigns target and distractor roles and freezes prediction-to-object correspondences to identify five error types and count their occurrences. Deletion, transformation, and insertion then construct outputs for all 32 repair subsets. TrackEval independently evaluates these outputs to obtain coalition scores and isolated gains. Finally,

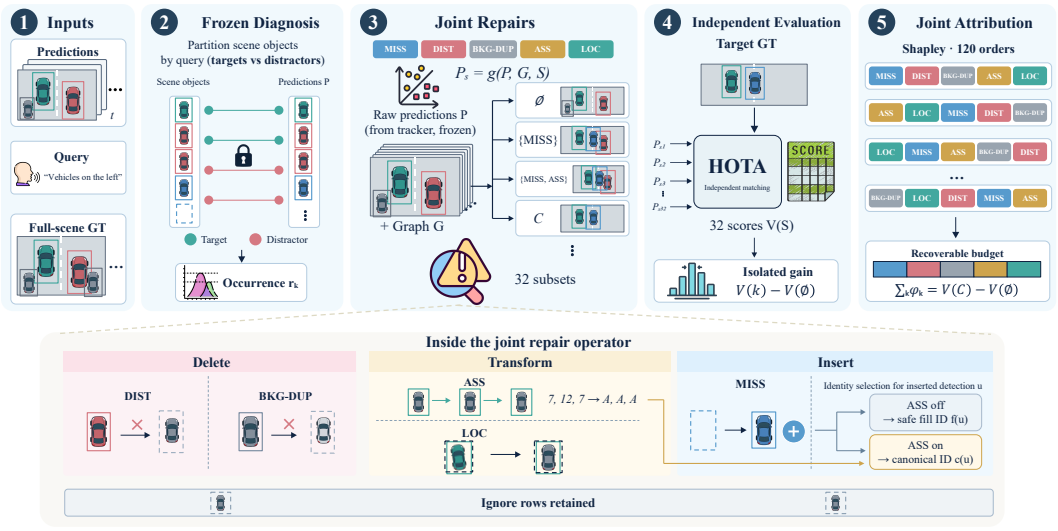



Figure 2: **Overview of RMOT-Compass.** The framework comprises five stages: **(1) Inputs:** provide frozen predictions, a language query, and full-scene ground truth. **(2) Frozen diagnosis:** assign target and distractor roles, fix prediction-to-object correspondences, and count observable errors. **(3) Joint repairs:** construct outputs for all 32 repair subsets. **(4) Independent evaluation:** rescore each output with fresh metric matching to obtain coalition values and isolated gains. **(5) Joint attribution:** average marginal gains over 120 repair orders to allocate the jointly recoverable gap. The lower panel details the repairs: DIST/BKG-DUP delete outputs, ASS/LOC transform retained predictions, and MISS inserts targets.

Shapley allocation averages marginal gains over all 120 repair orders, assigning each error type its share of the jointly recoverable gap.

### 3.2 QUERY-AWARE DIAGNOSTIC CORRESPONDENCE

**Query-conditioned scene roles.** For each frame in a video-query sequence, we assign evaluable scene objects a target or distractor role:

$$\rho_t(u; q) = \begin{cases} T, & u \in G_t^T(q), \\ D, & u \in G_t^D(q), \end{cases} \tag{1}$$

where $q$ is the language query, $t$ is the frame index, and $u$ is an evaluable scene object. The labels $T$ and $D$ denote target and distractor roles. Query annotations specify the target subset $G_t^T(q)$ of the full-scene object set $G_t^{\text{all}}$. The remaining objects form the distractor set $G_t^D(q) = G_t^{\text{all}} \setminus G_t^T(q)$. An optional ignore set $G_t^I$ is stored separately. Each object has a box and persistent identity. Each prediction $p \in P_t$ has a box $b_p$, track ID $i_p$, and score $s_p$.

**Joint correspondence to targets and distractors.** We jointly match predictions to targets and distractors without target priority, using the lexicographic objective

$$A_t^* = \operatorname*{lex\,arg\,max}_{A \in \mathfrak{A}_t(\tau_{\text{corr}})} \left( |A|, \sum_{(p,u) \in A} Q_{12}(\text{IoU}(b_p, b_u)) \right), \tag{2}$$

where $A_t^*$ is the selected matching, $\mathfrak{A}_t(\tau_{\text{corr}})$ contains feasible one-to-one matchings at correspondence threshold $\tau_{\text{corr}} = 0.05$, and $|A|$ is the number of matched pairs. Prediction $p$ and object $u$ have boxes $b_p$ and $b_u$. Their overlap is measured by IoU, and $Q_{12}(a) = \text{round}(10^{12} a)$ quantizes $a$.

**Frozen diagnostic correspondence.** We construct the diagnostic graph $\mathcal{G}$ once from the raw predictions $P = \bigcup_t P_t$. Query-conditioned roles and prediction-to-object correspondences remain fixed across all repair subsets, so repairs do not redefine the original error labels. This diagnostic correspondence is separate from the metric matching used for evaluation in Section 3.5. Matching details are in Appendix A.

Table 1: **Five observable error types and joint repairs.** Conditions use the frozen diagnostic graph. Repairs retain the evaluation ground truth.

| Error type | Observable condition | Repair when selected |
|---|---|---|
| MISS | Unmatched referred target | Insert GT box and coordinated ID |
| DIST | Matched annotated non-target | Remove prediction |
| BKG-DUP | Background or duplicate prediction | Remove prediction |
| ASS | Inconsistent target identity component | Assign canonical target ID |
| LOC | Target box IoU $< 1 - 10^{-6}$ | Replace with matched GT box |

### 3.3 FIVE OBSERVABLE ERROR TYPES

We define five observable error types on the frozen query-conditioned diagnostic graph. Table 1 summarizes their operational conditions and corresponding repairs.

**Selection errors.** Missing targets (MISS) are referred objects without a matched prediction. Distractor outputs (DIST) are predictions matched to annotated objects outside the query's target set. Background or duplicate outputs (BKG-DUP) are non-ignored residual predictions labeled as background or duplicates under the matching precedence in Appendix A.

**Trajectory errors.** Association errors (ASS) flag target-corresponding rows in inconsistent identity components, including fragmentation, identity merging, and contamination by non-target output. Localization errors (LOC) flag target-corresponding boxes that fail the near-exact IoU condition in Table 1. Association and localization errors may coexist on the same row.

**Event occurrence.** These types form the player set $\mathcal{C} = \{\text{MISS}, \text{DIST}, \text{BKG-DUP}, \text{ASS}, \text{LOC}\}$. We report their occurrence rates as

$$r_k = \frac{N_k}{D_k}, \qquad D_k = \begin{cases} N_{\text{target}}, & k = \text{MISS}, \\ N_{\text{pred}}, & k \in \{\text{DIST}, \text{BKG-DUP}\}, \\ N^T_{\text{match}}, & k \in \{\text{ASS}, \text{LOC}\}, \end{cases} \tag{3}$$

where $r_k$ is the occurrence rate of error type $k$, $N_k$ is its event count, and $D_k > 0$ is its reference population. The counts $N_{\text{target}}$, $N_{\text{pred}}$, and $N^T_{\text{match}}$ denote target detections, predictions, and target-matched predictions across evaluated queries.

### 3.4 QUERY-CONDITIONED JOINT REPAIRS

The joint operator in the lower panel of Fig. 2 combines deletion, transformation, and insertion to construct each repaired output:

$$P^S = g(P, \mathcal{G}, S) = \{T_S(p) : p \in P \setminus \mathcal{D}_S\} \cup \mathcal{M}_S, \tag{4}$$

where $P^S$ is the output for repair subset $S \subseteq \mathcal{C}$, $P$ is the raw prediction set, and $\mathcal{G}$ is its frozen diagnostic graph. The joint operator $g$ removes rows in $\mathcal{D}_S$, transforms retained rows through $T_S$, and adds target detections in $\mathcal{M}_S$.

**(1) Delete unwanted predictions.** Selecting DIST removes predictions matched to distractors, while BKG-DUP removes background or duplicate predictions. The frozen labels determine their union $\mathcal{D}_S$. Ignored rows remain in the output.

**(2) Transform retained predictions.** For retained target-matched predictions, $T_S$ assigns canonical target IDs across frames when ASS is selected and replaces boxes with matched ground-truth boxes when LOC is selected. Both corrections apply when selected together, while unselected fields retain their raw values.

**(3) Insert missing targets.** Selecting MISS inserts ground-truth boxes for unmatched targets into $\mathcal{M}_S$. Their track IDs depend on whether association repair is also selected:

$$i_S(u) = \begin{cases} c(u), & \text{ASS} \in S, \\ f(u), & \text{ASS} \notin S, \end{cases} \qquad \text{for insertions with MISS} \in S, \tag{5}$$

where $i_S(u)$ is the inserted target's identity. With ASS selected, the collision-free canonical ID $c(u)$ joins inserted and retained detections of target $u$. Otherwise, $f(u)$ reuses a safe raw ID or assigns a fresh one. Full identity rules are in Appendix A.

### 3.5 JOINT RECOVERABLE GAP AND ERROR ATTRIBUTION

**Independent evaluation and coalition values.** For each repaired output, TrackEval recomputes its matching against the query-target ground truth independently of the diagnostic graph. The resulting metric score defines the coalition value

$$V_M(S) = M\big(g(P, \mathcal{G}, S), G^T\big), \qquad S \subseteq \mathcal{C}, \tag{6}$$

where $V_M(S)$ is the value of repair subset $S$, $M$ is the tracking metric, $G^T$ is the query-target ground truth, and $g(P, \mathcal{G}, S)$ is the repaired output from Equation (4).

For each metric, the conditional marginal gain and joint recoverable gap are

$$\delta_k^M(S) = V_M(S \cup \{k\}) - V_M(S), \qquad J_M = V_M(\mathcal{C}) - V_M(\varnothing), \tag{7}$$

where $\delta_k^M(S)$ is the gain in metric $M$ from adding repair $k \notin S$ to subset $S$, and $J_M$ is the gain from all repairs in $\mathcal{C}$. The empty subset $\varnothing$ gives the raw output, with isolated gain $\Delta_k^{\text{iso}}(M) = \delta_k^M(\varnothing)$.

**Allocation across repair contexts.** We apply the standard Shapley value (Shapley et al., 1953) to this fixed game:

$$\phi_k(M) = \sum_{S \subseteq \mathcal{C} \setminus \{k\}} \frac{|S|!(n - |S| - 1)!}{n!}\, \delta_k^M(S), \qquad n = |\mathcal{C}| = 5, \tag{8}$$

where $\phi_k(M)$ is the allocation to error type $k$, $n$ is the number of error types, and $|S|$ is the size of a subset that excludes $k$. The factorial weight is the proportion of player orders in which exactly $S$ precedes $k$. The calculation averages marginals over all $5! = 120$ orders using the $2^5 = 32$ coalition scores, without rerunning the tracker.

**A complete metric budget.** Shapley efficiency assigns the entire jointly recoverable gain. We compare the all-repair output with an independently constructed ideal to obtain the full metric budget

$$\begin{aligned} M(P_{\text{ideal}}, G^T) &= V_M(\varnothing) + \sum_{k \in \mathcal{C}} \phi_k(M) + \text{RES}_M, \\ \text{RES}_M &= M(P_{\text{ideal}}, G^T) - V_M(\mathcal{C}), \end{aligned} \tag{9}$$

where $P_{\text{ideal}}$ is generated independently from target ground truth, $V_M(\varnothing)$ is the raw score, and $\sum_{k \in \mathcal{C}} \phi_k(M)$ is the allocated recovery. The residual $\text{RES}_M$ measures the gap between the all-repair and ideal scores.

Pairwise Shapley interaction indices (Grabisch & Roubens, 1999) describe average second-order gains within the same budget (Appendix A).

## 4 EXPERIMENTS

### 4.1 SETTINGS

**Datasets and methods.** We evaluate five representative open-source RMOT methods on the Refer-KITTI test set: end-to-end TransRMOT (Wu et al., 2023), TempRMOT (Zhang et al., 2024), and DKGTrack (Li et al., 2025), and two-stage iKUN (Du et al., 2024) and FlexHook (Li et al., 2026). TempRMOT and FlexHook also use the Refer-KITTI-V2 test set. Each method uses the authors' official dataset-specific weights without retraining (Appendix B).

**Evaluation.** For both datasets, we use a diagnostic IoU threshold of 0.05, a common GT registry format, and TrackEval to score all 32 repair subsets and an independently constructed ideal. We report exact global HOTA allocations on the 0–100 scale and event rates from Equation (3). On Refer-KITTI, we vary diagnostic IoU over 0.10, 0.30, and 0.50 with all other inputs and evaluation settings fixed (Appendix C). Additional metrics and query-macro uncertainty are in Appendix D.

Table 2: **Frozen RMOT methods evaluated on Refer-KITTI.** HOTA is evaluated on the test set. Every method passes the common input compatibility checks.

| Method | Prediction / candidate source | HOTA |
|---|---|---|
| TransRMOT | Released end-to-end checkpoint | 45.5 |
| TempRMOT | Released end-to-end checkpoint | 52.0 |
| DKGTrack | Released end-to-end checkpoint | 49.3 |
| iKUN | Frozen NeuralSORT tracklets | 44.3 |
| FlexHook | Frozen Temp-NeuralSORT candidates | 53.8 |

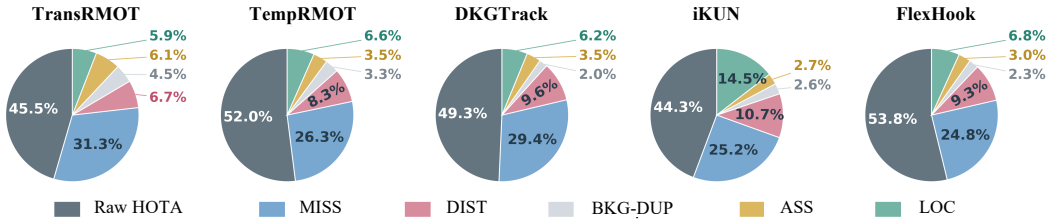



Figure 3: **Global HOTA budgets across five RMOT methods on Refer-KITTI.** Each pie combines raw HOTA with five Shapley allocations to total 100 points. Percentages refer to this complete budget, including the raw score, rather than the recoverable gap alone.

**Candidate perturbations.** With FlexHook's checkpoint and queries fixed, we apply target removal, distractor insertion, and ID fragmentation at doses 0.10 and 0.30, and box jitter at relative magnitudes 0.05 and 0.15 before inference. Event-rate changes use paired video-block bootstrap intervals from 10,000 replicates. Operator definitions are in Appendix E.

### 4.2 Error Attribution across RMOT Methods

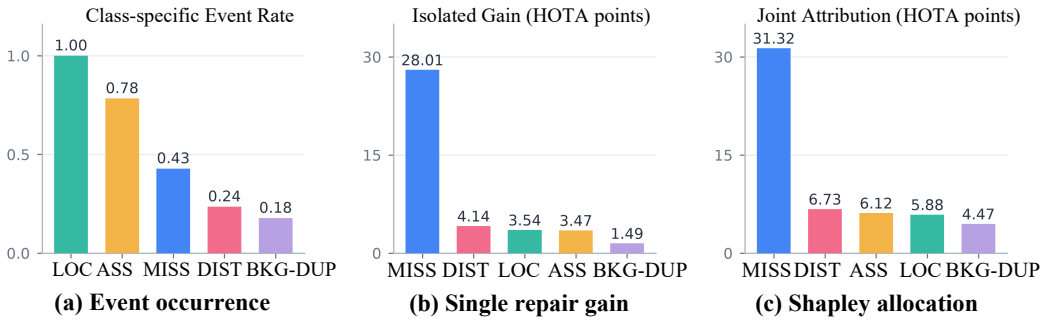
**Missing targets dominate joint recovery.** We report each method's HOTA on the Refer-KITTI test set together with its prediction or candidate source in Table 2. Fig. 3 decomposes the complete 100-point budget into raw HOTA and five error allocations. MISS contributes the largest error share for every method, at 24.8–31.3 HOTA points, and remains first at all four correspondence thresholds (Table 4 of Appendix C). These results show that missing targets dominate the recoverable error budget of all five evaluated methods.

**Higher HOTA can coexist with larger error contributions.** As shown in Fig. 3, FlexHook achieves 53.8 HOTA compared with TransRMOT's 45.5. Its MISS and ASS allocations are smaller, at 24.8 versus 31.3 and 3.0 versus 6.1 points. Its DIST allocation, however, is larger, at 9.3 versus 6.7 points, as is LOC, at 6.8 versus 5.9. Different inputs and training provenance preclude isolating architectural causes. We observe that higher HOTA can coexist with larger contributions from specific errors.

### 4.3 From Error Occurrence to Joint Attribution

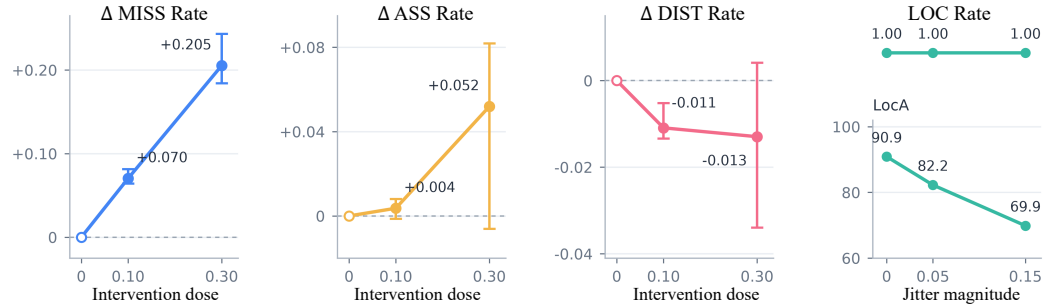
**Occurrence does not measure recoverable impact.** As illustrated in Fig. 4(a), LOC has the highest event rate, whereas MISS leads both HOTA-based rankings in Fig. 4(b) and Fig. 4(c). All five methods have LOC rate 1.00 under the near-exact box-equality condition, but their HOTA allocations span 5.88–14.49 points. Full profiles are in Table 6, Table 7 and Table 8 of Appendix D. These results show that counting box deviations alone does not capture their recoverable metric impact.

**Isolated gain and joint attribution yield different rankings.** Fig. 4(b) and Fig. 4(c) compare the same TransRMOT output. Repairing LOC alone gains 3.54 points, versus 3.47 for ASS. Joint attribution reverses their order: ASS receives 6.12 and LOC 5.88. Across the five methods, isolated gains sum to 7.43–13.88 fewer HOTA points than joint recovery (Table 9 of Appendix D). These results show that isolated gains give different rankings and underestimate the jointly recoverable gap.




**(a) Event occurrence** **(b) Single repair gain** **(c) Shapley allocation**

Figure 4: **Three diagnostic views of TransRMOT on Refer-KITTI.** (a) Event occurrence ranks error types by rates with class-specific denominators. (b) Isolated gain ranks them by the HOTA improvement from repairing each error alone. (c) Joint attribution ranks them by Shapley allocations of the recoverable gap. Panels (b) and (c) use HOTA points. LOC ranks above ASS in isolated gain but below it in joint attribution.




**(a) Target removal** **(b) ID fragmentation** **(c) Distractor insertion** **(d) Box jitter**

Figure 5: **FlexHook responses to candidate perturbations on Refer-KITTI.** (a) Target removal increases the MISS rate. (b) ID fragmentation increases the ASS point estimates, with intervals crossing zero. (c) Distractor insertion lowers the DIST point estimates as the prediction denominator expands. (d) Box jitter leaves the absolute LOC rate saturated (top) while reducing the auxiliary LocA score (bottom). Panels (a), (b) and (c) show changes from the unperturbed output. Open circles mark the zero-dose reference, and error bars are paired 95% video-block bootstrap intervals.

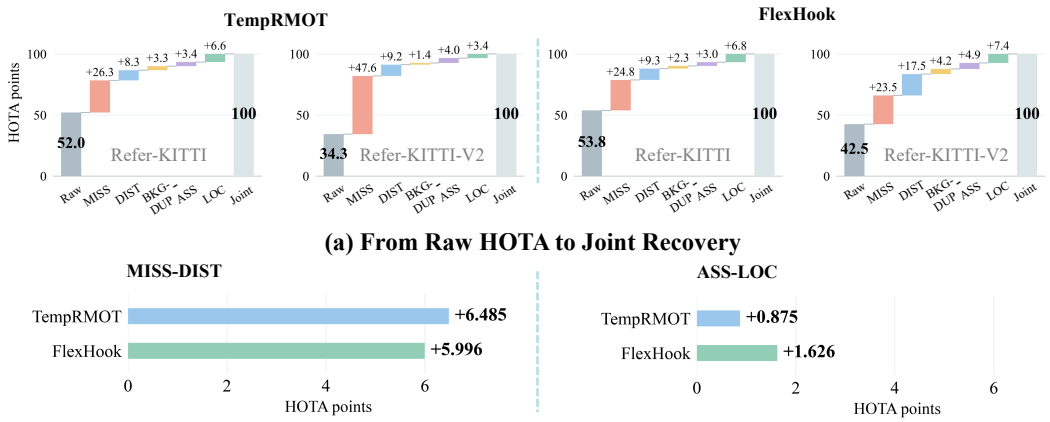
**Joint repairs reveal complementary gains.** Table 10 of Appendix D shows positive MISS–DIST and ASS–LOC interactions in all five methods. These positive interactions indicate that each repair becomes more beneficial, on average, when its paired repair is also applied.

### 4.4 CONTROLLED PERTURBATION ANALYSIS

**Target removal increases missing-target errors.** Fig. 5(a) shows MISS rate increases of $+0.070$ and $+0.205$ at doses 0.10 and 0.30, respectively. Target-associated candidate rows are removed before inference, while the checkpoint and language queries remain fixed. Both paired intervals exclude zero, with a higher-dose 95% CI of $[0.184, 0.243]$. These results confirm that the MISS rate tracks worsening target coverage at both perturbation doses.

**ID fragmentation increases association point estimates.** Fig. 5(b) reports ASS rate increases of $+0.004$ and $+0.052$ at the two doses. Fragmentation assigns a fresh ID to the second half of a selected candidate track, disrupting temporal continuity before inference. Both paired intervals cross zero despite the larger higher-dose estimate. The point estimates suggest an upward trend in association errors, although the intervals do not establish a systematic increase.

**Counts and rates can move in opposite directions.** As shown in Fig. 5(c), distractor insertion changes DIST rate by $-0.011$ and $-0.013$ at both doses. The higher-dose interval crosses zero. Table 13 of Appendix E reports increases in both distractor count and prediction count, with faster




**(a) From Raw HOTA to Joint Recovery**




**(b) Repair interactions on Refer-KITTI-V2**

Figure 6: **Cross-dataset HOTA budgets and repair interactions.** (a) Paired waterfalls show TempRMOT and FlexHook on Refer-KITTI and Refer-KITTI-V2. The sum of raw HOTA and the five Shapley allocations reaches 100. Intermediate heights accumulate allocations, not sequential-repair scores. (b) Horizontal bars show the prespecified interaction indices on Refer-KITTI-V2.

growth in the latter. We observe that a declining DIST rate can accompany more distractor errors when the prediction denominator grows faster.

**Geometric degradation can leave a binary event unchanged.** The curves in Fig. 5(d) show an unchanged LOC rate of 1.00 under box jitter, while LocA drops from 90.9 to 82.2 and 69.9. The perturbation tests evaluate event responses without computing new Shapley budgets. These results show why a saturated event rate needs continuous measures to reveal worsening localization quality.

### 4.5 PROTOCOL REUSE

We apply RMOT-Compass to TempRMOT and FlexHook on the Refer-KITTI-V2 dataset, retaining the error definitions, joint repair rules, and attribution procedure from the Refer-KITTI dataset.

**Method-specific error budgets.** As shown in Fig. 6(a), TempRMOT's HOTA decreases from 52.0 on Refer-KITTI to 34.3 on Refer-KITTI-V2, while its MISS allocation increases from 26.3 to 47.6 points. FlexHook's HOTA also decreases, from 53.8 to 42.5, but its MISS allocation changes from 24.8 to 23.5 points and its DIST allocation increases from 9.3 to 17.5. MISS remains dominant in both methods, while the shared definitions expose different changes in error composition behind lower aggregate scores.

**Complementary repairs and complete recovery.** Fig. 6(b) shows positive MISS–DIST indices of $+6.485$ for TempRMOT and $+5.996$ for FlexHook, alongside positive ASS–LOC indices of $+0.875$ and $+1.626$. As on Refer-KITTI, each repair gains more on average when its paired repair is present. Both all-repair outputs reach their independent ideal HOTA of 100 with zero residual (Fig. 6(a)). Full results are in Appendix G.

## 5 CONCLUSION

We introduce **RMOT-Compass**, a query-conditioned error attribution framework for referring multi-object tracking. It combines query-defined target and distractor roles, frozen diagnostic correspondences, and temporally coordinated joint repairs, then uses standard Shapley allocation to distribute the jointly recoverable HOTA gap among five observable error types. Evaluation of five RMOT methods on Refer-KITTI and two RMOT methods on Refer-KITTI-V2 identifies missing targets as the dominant contributor to this gap and shows that higher HOTA can coexist with larger contributions from specific errors. The differing rankings from event occurrence, isolated gain, and joint attribution highlight the need to interpret these diagnostic quantities separately. RMOT-Compass provides a quantitative basis for prioritizing observable tracking failures.

### AI USE STATEMENT

We use generative AI tools to assist with experimental implementation and language polishing. The authors review all AI-assisted work and take responsibility for the manuscript.

### REPRODUCIBILITY STATEMENT

Section 3 and Appendix A specify the diagnostic correspondence, error definitions, joint repairs, and attribution procedure. Section 4.1 and Appendix B describe the datasets, prediction sources, released checkpoints, and evaluation settings. All evaluated methods use released checkpoints without retraining. Appendix C reports implementation checks and correspondence-threshold sensitivity. Appendices D–G provide the statistical procedures and complete results for error attribution, controlled perturbations, query variation, and cross-dataset protocol reuse.

## REFERENCES

Jin Bai and Gregory D Hager. Cast: Evaluating multi-object trackers with context-aware switch and transfer scores. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7260–7268. IEEE, 2026.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP journal on image and video processing*, 2008(1):246309, 2008.

Daniel Bolya, Sean Foley, James Hays, and Judy Hoffman. Tide: A general toolbox for identifying object detection errors. In *European Conference on Computer Vision*, pp. 558–573. Springer, 2020.

Tzoulio Chamiti, Leandro Di Bella, Adrian Munteanu, and Nikos Deligiannis. Refergpt: Towards zero-shot referring multi-object tracking. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3849–3858. IEEE, 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2805–2813, 2026.

Sijia Chen, En Yu, and Wenbing Tao. Cross-view referring multi-object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2204–2211, 2025.

Yunhao Du, Cheng Lei, Zhicheng Zhao, and Fei Su. ikun: Speak to trackers without retraining. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19135–19144. IEEE, 2024.

Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of game theory*, 28(4):547–565, 1999.

Shukun Jia, Shiyu Hu, Yipei Wang, Ximeng Cheng, Yichao Cao, and Xiaobo Lu. Coal: Counterfactual and observation-enhanced alignment learning for discriminative referring multi-object tracking. *arXiv preprint arXiv:2605.14795*, 2026.

Guangyao Li, Siping Zhuang, Yajun Jian, Yan Yan, and Hanzi Wang. Language decoupling with fine-grained knowledge guidance for referring multi-object tracking. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23626–23635. IEEE, 2025.

Weize Li, Yunhao Du, Qixiang Yin, Zhicheng Zhao, and Fei Su. Rethinking two-stage referring-by-tracking in referring multi-object tracking: Make it strong again. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42549–42559, 2026.

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*, 124:103349, 2025.

Zijia Lu, Jingru Yi, Jue Wang, Yuxiao Chen, Junwen Chen, Xinyu Li, and Davide Modolo. Storm: End-to-end referring multi-object tracking in videos. *arXiv preprint arXiv:2604.10527*, 2026.

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.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

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.

Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pp. 17–35. Springer, 2016.

Lloyd S Shapley et al. A value for n-person games. 1953.

Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In *International conference on machine learning*, pp. 9269–9278. PMLR, 2020.

Dongming Wu, Wencheng Han, Tiancai Wang, Xingping Dong, Xiangyu Zhang, and Jianbing Shen. Referring multi-object tracking. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14633–14642. IEEE, 2023.

Yanqiu Yu, Zhifan Jin, Sijia Chen, Tongfei Chu, En Yu, Liman Liu, and Wenbing Tao. Rt-rmot: A dataset and framework for rgb-thermal referring multi-object tracking. *arXiv preprint arXiv:2602.22033*, 2026.

Xiangqun Zhang, Likai Wang, Zekun Qian, Ruize Han, and Wei Feng. Cd-rmot-bench: Benchmarking the cross-domain referring multi-object tracking. *arXiv preprint arXiv:2607.25239*, 2026.

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

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, pp. 10528–10536, 2025.

Zihan Zhou, Sijia Chen, Yanqiu Yu, En Yu, and Wenbing Tao. Ormot: A dataset and framework for omnidirectional referring multi-object tracking. *arXiv preprint arXiv:2603.05384*, 2026.

# A Formal Definitions and Joint Repairs

## A.1 Input Validation and Diagnostic Matching

**Valid inputs.** The diagnostic unit is a video-query sequence. Boxes use pixel coordinates $(x, y, w, h)$ and are not clipped before IoU computation. Input validation rejects invalid frame indices, non-finite coordinates or scores, non-positive sizes, and duplicate prediction identities within a frame. Target, distractor, and ignore sets are disjoint. Query-invalid frames have no referred targets and cannot produce MISS, ASS, or LOC events.

**Feasible matching.** The feasible one-to-one matching set in Equation (2) satisfies

$$\sum_u z_{pu} \le 1, \qquad \sum_p z_{pu} \le 1, \qquad z_{pu} = 0 \text{ if } \mathrm{IoU}(b_p, b_u) + 10^{-12} < \tau_{\mathrm{corr}}, \tag{10}$$

where $z_{pu} \in \{0, 1\}$ indicates a match between prediction $p$ and scene object $u$ in the same frame, $b_p$ and $b_u$ are their boxes, and $\tau_{\mathrm{corr}}$ is the correspondence threshold. The two sums limit each endpoint to one match, and $10^{-12}$ is the numerical tolerance.

The implementation encodes cardinality, integer-quantized IoU, and canonical edge order as lexicographically dominant integer weights and solves the assignment with the Hungarian algorithm. Prediction content keys include geometry, score, and the temporal track signature. Target membership does not affect the matching objective. Remaining indistinguishable-row ties do not create a semantic preference between identical trajectories.

**Residual labels.** A matched prediction inherits its endpoint's target/distractor role. An unmatched prediction is a duplicate if it overlaps an already matched evaluable object at $\tau_{\mathrm{corr}}$. Otherwise, it is ignored if it overlaps an ignore object at IoU 0.5, and background if neither condition holds. Duplicate and ignore anchors are selected by IoU and canonical GT keys. Unmatched referred objects become MISS events. These labels remain fixed across repair coalitions, with standard metric matching performed independently afterward.

## A.2 Identity, Localization, and Denominators

**Association components.** Across a sequence, we link each raw prediction identity to its matched target identities in a bipartite graph. An identity component is association-inconsistent when

$$a(K) = \mathbf{1}\{|\mathcal{I}_K| > 1 \ \vee \ |\mathcal{U}_K| > 1 \ \vee \ |\mathcal{C}_K| > 0\}, \tag{11}$$

where $K$ is a connected component, $\mathcal{I}_K$ and $\mathcal{U}_K$ are its prediction and target identity sets, and $\mathcal{C}_K$ contains prediction identities also used by distractor, duplicate, or background rows. The indicator $\mathbf{1}\{\cdot\}$ equals one when its condition holds, and $a(K)$ flags an inconsistent component.

Every target-corresponding row in such a component contributes to the ASS count. This covers fragmentation, merged or switched identities, and contamination by non-target output. Ignore rows do not cause contamination. Prediction IDs need not numerically equal GT IDs, and a missing frame alone is not an association event.

**Localization and semantic co-occurrence.** We define localization severity and events by

$$\ell(p, u) = 1 - \mathrm{IoU}(b_p, b_u), \qquad e_{\mathrm{Loc}}(p, u) = \mathbf{1}\{\ell(p, u) > 10^{-6}\}, \tag{12}$$

where $(p, u)$ is a prediction-to-target correspondence, $\ell(p, u)$ is its localization severity, and $e_{\mathrm{Loc}}(p, u)$ is the binary localization event. Here, IoU compares boxes $b_p$ and $b_u$, and $\mathbf{1}\{\cdot\}$ denotes an indicator.

An auxiliary severe-error tag uses IoU below 0.5. Association and localization may label the same row. SEM-SUB marks a query-frame containing at least one MISS and one DIST, without identifying a particular substituted object pair. Event-rate denominators are given in Equation (3), and SEM-SUB uses query-frames. Neither SEM-SUB nor the ideal residual enters the five-player game.

### A.3 Composition, Identities, and Provenance

**Transforming retained rows.** Each retained prediction receives the transformed box and identity

$$b_p^S = \begin{cases} b_{u(p)}, & p \text{ is target-matched, } e_{\mathrm{Loc}}(p, u(p)) = 1, \ \textsc{Loc} \in S, \\ b_p, & \text{otherwise,} \end{cases}$$
$$i_p^S = \begin{cases} c(u(p)), & p \text{ is target-matched, } \textsc{Ass} \in S, \\ i_p, & \text{otherwise.} \end{cases} \tag{13}$$

where $b_p^S$ and $i_p^S$ are the repaired box and identity of prediction $p$, $u(p)$ is its frozen target endpoint, and $b_{u(p)}$ is that target's ground-truth box. The function $c$ assigns a canonical target ID, $e_{\mathrm{Loc}}$ flags localization error, and $S$ is the selected repair subset. The full row transformation is $T_S(p) = (t, b_p^S, i_p^S, s_p)$, preserving frame $t$ and score $s_p$.

Deletion uses only frozen spatial labels: Dist removes distractor rows, and Bkg-Dup removes background or duplicate rows. Ignore rows are retained. With Miss, insertions use missing targets' GT boxes, score 1, and identities from Equation (5).

**Constructing safe identities.** Canonical target IDs are unique and lie outside the raw prediction-ID range. A target-linked raw identity can be reused for insertions only if

$$F_{\mathrm{miss}}(u) \cap \big(F_{\mathrm{raw}}(i) \cup F_{\mathrm{reserved}}(i)\big) = \varnothing, \tag{14}$$

where $u$ is the target and $i$ a candidate raw identity. The sets $F_{\mathrm{miss}}(u)$, $F_{\mathrm{raw}}(i)$, and $F_{\mathrm{reserved}}(i)$ contain insertion frames, frames occupied by $i$, and frames reserved by earlier fills, respectively.

Eligible IDs are ranked by target correspondence count, total IoU, and track signature. Targets are processed in canonical order, with insertion frames reserved after each choice. If no candidate is safe, $f(u) = c(u)$. The fill plan is computed from raw rows and reused across coalitions, so deleting another row does not silently change the fill identity. Activating Ass maps both matched and inserted target rows to $c(u)$.

**Derived-output contract.** Each derived row records its source, graph, repair subset, and modifications. Reapplying the same subset to compatible derived rows returns the existing output. Mixing raw and derived rows, changing the graph or subset of a derived input, and using an unmatched source record are rejected. Other coalitions are constructed from raw predictions. The ideal output is generated independently from query-target GT, not copied from the all-repair output.

### A.4 Allocation, Interaction, and Residual

**Permutation form and efficiency.** The permutation form of Equation (8) makes the averaging over repair orders explicit:

$$\phi_k(M) = \frac{1}{n!} \sum_{\pi \in \Pi(\mathcal{C})} \delta_k^M(\mathrm{Pre}_\pi(k)), \tag{15}$$

where $\Pi(\mathcal{C})$ contains all $n!$ permutations of the $n$ error types, $\pi$ is one permutation, and $\mathrm{Pre}_\pi(k)$ is the set preceding error type $k$. The term $\delta_k^M$ is its conditional metric gain, and $\phi_k(M)$ is its allocation.

For a fixed predecessor set $S$, exactly $|S|!(n - |S| - 1)!$ permutations place $S$ before $k$, giving the weights in Equation (8). Along each permutation, marginal gains telescope:

$$\sum_{k \in \mathcal{C}} \delta_k^M(\mathrm{Pre}_\pi(k)) = V_M(\mathcal{C}) - V_M(\varnothing) = J_M, \tag{16}$$

where $V_M(\mathcal{C})$ and $V_M(\varnothing)$ are the all-repair and raw scores, and $J_M$ is their difference. Averaging over permutations gives $\sum_k \phi_k(M) = J_M$, the standard Shapley efficiency identity used in Equation (9).

**Pairwise interaction.** The joint effect of two repairs is captured by the second difference

$$\Delta_{ij} V(S) = V(S \cup \{i, j\}) - V(S \cup \{i\}) - V(S \cup \{j\}) + V(S), \tag{17}$$

where $i$ and $j$ are distinct error types outside subset $S$, $V$ is the coalition value with metric subscript omitted, and $\Delta_{ij} V(S)$ compares their joint gain with the sum of their separate gains in that context.

The reported Shapley interaction index (Grabisch & Roubens, 1999) is

$$I_{ij}(V) = \sum_{S \subseteq \mathcal{C} \setminus \{i,j\}} \frac{|S|!(n - |S| - 2)!}{(n-1)!} \Delta_{ij} V(S), \tag{18}$$

where $I_{ij}(V)$ is the pairwise interaction index, $n = |\mathcal{C}|$ is the number of error types, and the factorial weights average $\Delta_{ij}V(S)$ over subsets excluding $i$ and $j$. Its sign describes average complementarity, although individual contexts can have different signs. The index has metric units, and its contribution is already distributed within the Shapley allocations.

**Non-additivity versus incomplete recovery.** We measure the difference between joint recovery and the sum of isolated gains as

$$\Gamma_M = J_M - \sum_{k \in \mathcal{C}} \Delta_k^{\text{iso}}(M), \tag{19}$$

where $M$ is the tracking metric, $\mathcal{C}$ is the set of five error types, and $k$ indexes one error type. For a repair subset $S$, $V_M(S)$ is the metric score after applying its joint repair, with $\varnothing$ leaving the output unrepaired. The joint recoverable gap is $J_M = V_M(\mathcal{C}) - V_M(\varnothing)$, and the isolated gain is $\Delta_k^{\text{iso}}(M) = V_M(\{k\}) - V_M(\varnothing)$. Their difference, $\Gamma_M$, measures non-additivity relative to the raw output and can include interactions of more than two repairs.

The separate residual $\text{RES}_M$ compares the all-repair output with the independently constructed ideal. A zero residual establishes complete recovery under the chosen registry and repair family.

# B Data and Method Provenance

**Data records.** The GT registry records dataset, split, video, query, frame validity, object identity, target/distractor role, box, category where informative, and ignore status. Prediction records retain the method, checkpoint, query, frame, identity, box, score, candidate source, and source-row provenance. Method cards additionally record the environment, inference command, normalization rule, and evaluator revision.

**Method inputs.** TransRMOT, TempRMOT, and DKGTrack use released end-to-end checkpoints and frozen prediction directories. iKUN uses NeuralSORT car/pedestrian tracklets, the released cascade-attention checkpoint, and similarity calibration. FlexHook uses Temp-NeuralSORT-kitti1 candidates. A common evaluator standardizes the scored outputs, while candidate sources and training conditions remain properties of each method.

**Normalization.** The central loader sorts rows by identity, box, and remaining columns. Without this normalization, CLEAR tie matching yields native-order differences of up to four ID switches in three methods, with unchanged HOTA, IDF1, and fragmentation. All reported tables use the corrected central values.

# C Validation and Sensitivity

## C.1 Validation of the Diagnostic Contract

Table 3 separates transformation checks, repeated evaluation, numerical accounting, and correspondence sensitivity. Each check addresses a distinct part of the diagnostic contract.

**Fixture coverage.** The 11 fixtures cover perfect output, missing targets, distractors, background, duplicates, identity switches or fragmentation, and box jitter. Combined-error cases include Miss with Dist, Ass with Loc, and all five errors. The remaining fixture contains perfect output with an empty frame. An empty-video case is tested separately. All 24 recorded tests pass with the production HOTA evaluator. Each fixture evaluates all 32 coalitions, recovers the declared error players, and reaches an independently generated ideal after all repairs. These controlled inputs test the declared transformations and their composition.

**Numerical audit.** We measure allocation efficiency error by

$$\epsilon_{\text{eff}}^{(m)} = \left| \sum_{k \in \mathcal{C}} \phi_k^{(m)} - \left( V^{(m)}(\mathcal{C}) - V^{(m)}(\varnothing) \right) \right|, \tag{20}$$

Table 3: **Validation of the diagnostic contract.** Repeatability is measured at saved-summary precision. Invariance results retain the tested toy, graph, and baseline scopes.

| Check | Scope | Result |
|---|---|---|
| Fixture recovery | 11 fixtures, 32 coalitions each | 24 tests pass |
| Repeatability | 160 summaries, 39 fields | Max. difference 0, tolerance $10^{-12}$ |
| Row / ID invariance | Toy, graph, and baseline tests | Pass within tested scopes |
| Efficiency | Five global HOTA games | Error $< 10^{-9}$ |
| Marginal edges | 80 per primary HOTA game | 0 of 400 negative |
| Ideal recovery | All-repair vs. independent ideal | HOTA 100, residual 0 |
| Nearby threshold | Event counts: IoU 0.05 vs. 0.10 | Top-three 3/3, rank correlation 0.9–1.0 |

Table 4: **HOTA attribution under correspondence-threshold changes on Refer-KITTI.** Each row evaluates all 32 repair subsets from the same raw predictions and GT. Only the diagnostic correspondence threshold changes, while metric evaluation remains fixed. Allocations are HOTA points, and Largest identifies the highest allocation.

| Method | IoU | MISS | DIST | BKG-DUP | ASS | LOC | Largest |
|---|---|---|---|---|---|---|---|
| TransRMOT | 0.05 | 31.325 | 6.730 | 4.465 | 6.120 | 5.884 | MISS |
| TransRMOT | 0.10 | 31.551 | 6.670 | 4.607 | 5.999 | 5.696 | MISS |
| TransRMOT | 0.30 | 31.796 | 6.377 | 5.056 | 5.916 | 5.379 | MISS |
| TransRMOT | 0.50 | 31.720 | 6.260 | 5.223 | 6.296 | 5.024 | MISS |
| TempRMOT | 0.05 | 26.335 | 8.336 | 3.280 | 3.404 | 6.632 | MISS |
| TempRMOT | 0.10 | 26.450 | 8.305 | 3.368 | 3.382 | 6.482 | MISS |
| TempRMOT | 0.30 | 26.640 | 8.164 | 3.728 | 3.328 | 6.128 | MISS |
| TempRMOT | 0.50 | 26.536 | 7.976 | 4.030 | 3.928 | 5.519 | MISS |
| DKGTrack | 0.05 | 29.379 | 9.613 | 2.016 | 3.446 | 6.244 | MISS |
| DKGTrack | 0.10 | 29.500 | 9.517 | 2.205 | 3.407 | 6.069 | MISS |
| DKGTrack | 0.30 | 29.513 | 9.278 | 2.594 | 3.574 | 5.740 | MISS |
| DKGTrack | 0.50 | 29.558 | 9.155 | 2.790 | 3.838 | 5.356 | MISS |
| iKUN | 0.05 | 25.154 | 10.735 | 2.640 | 2.642 | 14.493 | MISS |
| iKUN | 0.10 | 25.098 | 10.720 | 2.710 | 2.749 | 14.386 | MISS |
| iKUN | 0.30 | 25.149 | 10.512 | 3.119 | 3.260 | 13.624 | MISS |
| iKUN | 0.50 | 25.789 | 10.013 | 3.877 | 4.062 | 11.922 | MISS |
| FlexHook | 0.05 | 24.837 | 9.275 | 2.284 | 2.993 | 6.787 | MISS |
| FlexHook | 0.10 | 24.880 | 9.257 | 2.337 | 2.997 | 6.705 | MISS |
| FlexHook | 0.30 | 24.538 | 9.177 | 2.589 | 3.690 | 6.182 | MISS |
| FlexHook | 0.50 | 24.496 | 9.017 | 2.882 | 4.124 | 5.657 | MISS |

where $m$ indexes a frozen method, $\phi_k^{(m)}$ is its HOTA allocation to error type $k$, and $V^{(m)}$ is its HOTA coalition value. The absolute discrepancy $\epsilon_{\text{eff}}^{(m)}$ measures numerical efficiency error.

Across the five frozen methods, all 160 coalition summaries reproduce in all 39 numeric fields, with maximum difference zero under a $10^{-12}$ tolerance. All five primary HOTA games satisfy $\epsilon_{\text{eff}}^{(m)} < 10^{-9}$, with a recorded error of $7.11 \times 10^{-15}$ for TransRMOT. All-repair outputs reach their independent ideals with zero residual, a check separate from the repeated evaluations. None of the 400 primary-HOTA marginal edges is negative, although allocations for other metrics can be negative.

**Invariance checks.** Row-order and bijective-ID tests cover toy coalition values, semantic graph labels, and baseline row-order comparisons. They do not constitute transformed-output reevaluation of every coalition of each evaluated method.

### C.2 CORRESPONDENCE SENSITIVITY

**Fixed-output protocol.** We vary correspondence IoU over $\{0.05, 0.10, 0.30, 0.50\}$ for all five RMOT methods on Refer-KITTI. Each threshold uses its own graph, frozen before repair, while predictions,

Table 5: **Event-count sensitivity across independently built diagnostic graphs.** These are event counts, not Shapley allocations. The recorded nearby-threshold rank audit uses these counts.

| Method | IoU | MISS | DIST | BKG-DUP | ASS | LOC |
|---|---|---|---|---|---|---|
| TransRMOT | 0.05 | 28,479 | 15,237 | 11,495 | 29,757 | 37,982 |
| TransRMOT | 0.10 | 28,650 | 14,944 | 11,959 | 28,727 | 37,811 |
| TransRMOT | 0.30 | 28,970 | 13,937 | 13,286 | 27,962 | 37,491 |
| TransRMOT | 0.50 | 29,276 | 13,705 | 13,824 | 28,310 | 37,185 |
| TempRMOT | 0.05 | 25,055 | 17,410 | 6,594 | 29,666 | 41,406 |
| TempRMOT | 0.10 | 25,182 | 17,300 | 6,831 | 28,177 | 41,279 |
| TempRMOT | 0.30 | 25,379 | 16,720 | 7,608 | 27,041 | 41,082 |
| TempRMOT | 0.50 | 25,886 | 16,386 | 8,449 | 27,645 | 40,575 |
| DKGTrack | 0.05 | 28,372 | 20,864 | 4,402 | 28,793 | 38,089 |
| DKGTrack | 0.10 | 28,514 | 20,542 | 4,866 | 28,096 | 37,947 |
| DKGTrack | 0.30 | 28,750 | 19,826 | 5,818 | 26,995 | 37,711 |
| DKGTrack | 0.50 | 29,076 | 19,598 | 6,372 | 27,330 | 37,385 |
| iKUN | 0.05 | 25,752 | 24,408 | 5,001 | 27,335 | 40,709 |
| iKUN | 0.10 | 25,842 | 24,370 | 5,129 | 27,185 | 40,619 |
| iKUN | 0.30 | 26,256 | 23,894 | 6,019 | 29,107 | 40,205 |
| iKUN | 0.50 | 27,785 | 22,973 | 8,469 | 29,056 | 38,676 |
| FlexHook | 0.05 | 23,028 | 18,894 | 3,993 | 29,891 | 43,433 |
| FlexHook | 0.10 | 23,092 | 18,858 | 4,093 | 29,081 | 43,369 |
| FlexHook | 0.30 | 23,388 | 18,651 | 4,596 | 29,132 | 43,073 |
| FlexHook | 0.50 | 23,838 | 18,389 | 5,308 | 29,816 | 42,623 |

query targets, full-scene GT, and repair rules remain fixed. We evaluate all 32 repair subsets per method and threshold, for 640 coalition evaluations, and recompute exact Shapley allocations. TrackEval's metric-matching thresholds remain fixed.

**HOTA attribution.** Table 4 reports the five error allocations for every method and threshold. MISS ranks first in all 20 cases, with allocations from 24.496 to 31.796 HOTA points. Each method's raw HOTA remains unchanged, and all-repair HOTA reaches 100. The maximum absolute discrepancy between total attribution and the jointly recoverable HOTA gap is $1.42\times10^{-14}$. The primary-threshold allocations reproduce the reported results. Individual budgets still depend on the correspondence: iKUN's LOC allocation decreases from 14.493 to 11.922 points between thresholds 0.05 and 0.50, and TempRMOT's BKG-DUP overtakes ASS at 0.30.

**Event counts.** Table 5 reports counts from the same diagnostic graphs. Raising correspondence IoU from 0.05 to 0.10 preserves the top three event-count categories in every method, with rank correlations of 0.9 for DKGTrack and 1.0 for the others.

# D FULL DIAGNOSTIC RESULTS

## D.1 GLOBAL METRICS, ALLOCATIONS, AND EVENT RATES

**Raw tracking quality.** Table 6 reports raw HOTA alongside detection, association, localization, and identity metrics. These scores are the unrepaired reference for each diagnostic budget.

**Metric-specific recovery.** As shown in Table 7, an error type's allocation depends on the metric being explained. Allocations are computed from saved global coalition summaries before display rounding. All primary HOTA residuals are zero, so the raw score and five allocations sum to the independently evaluated ideal.

**Occurrence and reference populations.** Table 8 reports the event rates and defines their reference populations. The rates describe different populations and cannot be added or interpreted as shares of the recoverable score gap.

Table 6: **Central Refer-KITTI metrics for all frozen methods.** Each method uses 150 queries and 66,461 target detections.

| Method | HOTA | DetA | AssA | LocA | IDF1 | IDSW | Frag |
|---|---|---|---|---|---|---|---|
| TransRMOT | 45.476 | 35.813 | 57.969 | 90.240 | 51.574 | 583 | 1551 |
| TempRMOT | 52.012 | 40.561 | 66.861 | 90.439 | 59.391 | 264 | 1186 |
| DKGTrack | 49.302 | 36.849 | 66.080 | 90.393 | 55.053 | 274 | 1139 |
| iKUN | 44.337 | 31.960 | 62.036 | 81.045 | 55.237 | 174 | 900 |
| FlexHook | 53.824 | 43.354 | 66.916 | 90.872 | 62.750 | 183 | 966 |

Table 7: **Global metric Shapley allocations and ideal residuals on Refer-KITTI.** Values are points on each metric's 0–100 scale. Negative allocations are retained.

| Metric | Method | MISS | DIST | BKG-DUP | ASS | LOC | RES |
|---|---|---|---|---|---|---|---|
| HOTA | TransRMOT | 31.325 | 6.730 | 4.465 | 6.120 | 5.884 | 0.000 |
| HOTA | TempRMOT | 26.335 | 8.336 | 3.280 | 3.404 | 6.632 | 0.000 |
| HOTA | DKGTrack | 29.379 | 9.613 | 2.016 | 3.446 | 6.244 | 0.000 |
| HOTA | iKUN | 25.154 | 10.735 | 2.640 | 2.642 | 14.493 | 0.000 |
| HOTA | FlexHook | 24.837 | 9.275 | 2.284 | 2.993 | 6.787 | 0.000 |
| DetA | TransRMOT | 35.541 | 12.094 | 9.359 | 0.215 | 6.978 | 0.000 |
| DetA | TempRMOT | 31.720 | 14.446 | 5.757 | 0.081 | 7.435 | 0.000 |
| DetA | DKGTrack | 35.827 | 16.666 | 3.757 | 0.092 | 6.810 | 0.000 |
| DetA | iKUN | 31.240 | 18.004 | 3.988 | 0.127 | 14.681 | 0.000 |
| DetA | FlexHook | 29.410 | 16.300 | 3.635 | 0.082 | 7.220 | 0.000 |
| AssA | TransRMOT | 25.755 | 0.216 | -1.468 | 13.177 | 4.351 | 0.000 |
| AssA | TempRMOT | 19.301 | 0.732 | 0.267 | 7.395 | 5.443 | 0.000 |
| AssA | DKGTrack | 20.564 | 0.502 | -0.159 | 7.657 | 5.356 | 0.000 |
| AssA | iKUN | 16.760 | 0.811 | 0.968 | 6.022 | 13.403 | 0.000 |
| AssA | FlexHook | 19.096 | 0.674 | 0.709 | 6.451 | 6.155 | 0.000 |
| LocA | TransRMOT | 2.250 | 0.096 | 0.040 | 0.047 | 7.328 | 0.000 |
| LocA | TempRMOT | 1.957 | 0.088 | 0.007 | 0.010 | 7.500 | 0.000 |
| LocA | DKGTrack | 2.217 | 0.114 | 0.010 | 0.009 | 7.258 | 0.000 |
| LocA | iKUN | 4.184 | 0.041 | 0.001 | 0.009 | 14.721 | 0.000 |
| LocA | FlexHook | 1.704 | 0.070 | 0.000 | 0.021 | 7.333 | 0.000 |

Table 8: **Primary-graph event rates (IoU 0.05).** The denominators are target GT detections for MISS, predictions for DIST/BKG-DUP, and target-matched predictions for ASS/LOC. Rates are not additive.

| Method | MISS | DIST | BKG-DUP | ASS | LOC |
|---|---|---|---|---|---|
| TransRMOT | 0.4285 | 0.2355 | 0.1776 | 0.7835 | 1.0000 |
| TempRMOT | 0.3770 | 0.2662 | 0.1008 | 0.7165 | 1.0000 |
| DKGTrack | 0.4269 | 0.3293 | 0.0695 | 0.7559 | 1.0000 |
| iKUN | 0.3875 | 0.3481 | 0.0713 | 0.6715 | 1.0000 |
| FlexHook | 0.3465 | 0.2849 | 0.0602 | 0.6882 | 1.0000 |

### D.2 ISOLATED GAINS AND PERMUTATION MARGINALS

The 32 coalition values determine all 120 repair permutations through Equation (15). Table 9 compares their joint gap with the sum of isolated gains and reports the range of DIST marginals. These ranges describe repair contexts, not sampling uncertainty. Permutation averages agree with Shapley allocations, and MISS is the largest marginal in every permutation for each method. The dominant category is stable even though the amount of credit changes across orders.

Table 9: **Isolated gains and permutation marginals from the saved five-player HOTA games.** All values are HOTA points. Ranges span repair contexts, and the 120 permutations are not independent samples.

| Method | Joint gap | Isolated sum | Difference | DIST range | DIST Shapley |
|---|---|---|---|---|---|
| TransRMOT | 54.524 | 40.645 | 13.879 | 4.139–9.806 | 6.730 |
| TempRMOT | 47.988 | 38.443 | 9.545 | 5.800–10.982 | 8.336 |
| DKGTrack | 50.698 | 41.075 | 9.623 | 6.778–12.760 | 9.613 |
| iKUN | 55.663 | 48.238 | 7.425 | 6.990–14.565 | 10.735 |
| FlexHook | 46.176 | 38.423 | 7.753 | 6.785–11.759 | 9.275 |

Table 10: **All pairwise HOTA interaction indices on Refer-KITTI.** MISS–DIST and ASS–LOC are prespecified pairs. Values are descriptive HOTA points.

| Pair | TransRMOT | TempRMOT | DKGTrack | iKUN | FlexHook |
|---|---|---|---|---|---|
| MISS–DIST | **+2.999** | **+2.915** | **+3.814** | **+3.574** | **+2.975** |
| MISS–BKG-DUP | +2.385 | +1.141 | +0.847 | +0.688 | +0.666 |
| MISS–ASS | +0.758 | +1.665 | +1.148 | +0.697 | +1.781 |
| MISS–LOC | +0.797 | +0.467 | +0.288 | -2.131 | +0.278 |
| DIST–BKG-DUP | +1.434 | +1.053 | +0.813 | +1.042 | +0.735 |
| DIST–ASS | +0.439 | -0.076 | +0.166 | -0.209 | -0.132 |
| DIST–LOC | +0.794 | +1.041 | +1.129 | +2.754 | +1.129 |
| BKG-DUP–ASS | +1.585 | -0.142 | +0.195 | -0.759 | -0.587 |
| BKG-DUP–LOC | +0.773 | +0.470 | +0.289 | +0.736 | +0.292 |
| ASS–LOC | +2.583 | +1.748 | +1.707 | +1.617 | +1.098 |

### D.3 INTERACTIONS AND QUERY-MACRO UNCERTAINTY

**Interactions and cases.** Table 10 reports all ten HOTA interaction indices, including the prespecified MISS–DIST and ASS–LOC pairs. These are point estimates within the fixed game. The qualitative case inventory contains 25 method-by-player groups, each with an observed failure and its successful frozen repair, for 50 records. Fig. 1(a) illustrates query roles, while the quantitative panels use saved aggregate coalitions.

**Two estimands.** Averaging query-level allocations gives the query-macro estimand

$$\overline{\phi}_k = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \phi_{k,q}, \qquad \overline{\phi}_k \text{ need not equal } \phi_k^{\text{global}}, \tag{21}$$

where $\mathcal{Q}$ is the evaluated query set, $q$ indexes a query, $\phi_{k,q}$ is its allocation to error type $k$, and $\overline{\phi}_k$ is their mean. The global allocation $\phi_k^{\text{global}}$ uses the globally aggregated metric game.

The two quantities differ because global HOTA is computed by the evaluator's nonlinear aggregation. For example, TransRMOT's global MISS allocation is 31.324533 points, whereas its query-macro estimate is 43.288115, with 95% interval $[32.499364, 48.591020]$.

Table 11 uses 10,000 video-block bootstrap replicates with a fixed base seed. For each method, the implementation initializes Python's pseudorandom generator with the base seed plus the sum of character code points in its exact method name. Each replicate samples videos with replacement and includes all queries in each sampled video.

## E CONTROLLED CANDIDATE TESTS

### E.1 CONDITIONS AND CORRECTED ASSOCIATIONS

**Fixed components and perturbation units.** FlexHook's checkpoint, query list, and evaluator are fixed. Candidate tracks are modified before inference, and each condition receives its own output graph and event evaluation. The primary endpoints are MISS, DIST, ASS, and binary LOC for

Table 11: **Mean query-level HOTA Shapley allocations with 95% video-block intervals.** Three videos, 150 queries, 10,000 replicates, fixed base seed. These differ from global HOTA allocations.

| Method | Player | Query-macro points | 95% CI |
|---|---|---|---|
| TransRMOT | MISS | 43.288 | [32.499, 48.591] |
| TransRMOT | DIST | 6.980 | [5.484, 9.914] |
| TransRMOT | BKG-DUP | 4.946 | [1.054, 13.968] |
| TransRMOT | ASS | 3.879 | [2.880, 4.971] |
| TransRMOT | LOC | 5.717 | [3.715, 10.931] |
| TempRMOT | MISS | 35.412 | [31.658, 37.420] |
| TempRMOT | DIST | 8.724 | [7.339, 9.838] |
| TempRMOT | BKG-DUP | 3.628 | [0.979, 10.592] |
| TempRMOT | ASS | 2.674 | [1.325, 4.002] |
| TempRMOT | LOC | 6.565 | [4.593, 10.670] |
| DKGTrack | MISS | 43.264 | [37.334, 46.584] |
| DKGTrack | DIST | 8.496 | [6.854, 9.230] |
| DKGTrack | BKG-DUP | 2.234 | [0.998, 5.565] |
| DKGTrack | ASS | 2.248 | [1.314, 2.765] |
| DKGTrack | LOC | 5.595 | [4.253, 8.280] |
| iKUN | MISS | 52.764 | [46.752, 68.905] |
| iKUN | DIST | 8.953 | [5.818, 12.412] |
| iKUN | BKG-DUP | 2.346 | [1.400, 3.928] |
| iKUN | ASS | 1.586 | [0.617, 2.016] |
| iKUN | LOC | 9.015 | [7.419, 10.761] |
| FlexHook | MISS | 30.657 | [24.068, 33.458] |
| FlexHook | DIST | 10.852 | [7.553, 15.016] |
| FlexHook | BKG-DUP | 3.115 | [0.279, 9.951] |
| FlexHook | ASS | 1.872 | [1.441, 2.086] |
| FlexHook | LOC | 6.180 | [5.081, 7.907] |

target removal, distractor insertion, fragmentation, and jitter, respectively. These experiments do not compute corruption-specific 32-coalition allocations.

**Perturbation operators.** At dose $d \in \{0.10, 0.30\}$, target removal independently selects eligible target-associated rows with probability $d$. Distractor insertion selects eligible annotated distractor tracks with probability $d$ and adds their GT boxes under fresh candidate identities. Eligibility uses framewise unions of query roles in the frozen query list. Fragmentation selects tracks of at least two rows and assigns a fresh ID to the temporally ordered second half. Box jitter changes each candidate box according to

$$\begin{aligned} x' &= \max(0, x + dw\epsilon_x), \quad y' = \max(0, y + dh\epsilon_y), \\ w' &= \max(1, w + dw\epsilon_w), \quad h' = \max(1, h + dh\epsilon_h), \end{aligned} \tag{22}$$

where $(x, y, w, h)$ is the original box position, width, and height, and primes denote perturbed coordinates. The magnitude is $d \in \{0.05, 0.15\}$, and $\epsilon_x, \epsilon_y, \epsilon_w, \epsilon_h \sim \mathcal{N}(0, 1)$ are independent standard normal draws.

The dose parameters describe different operations: removal and insertion proportions, fragmentation probability, and box-jitter magnitude.

**Corrected candidate associations.** All 38,275 scene rows encode category zero, so earlier assumption of separate car/pedestrian labels gives incorrect candidate associations. The active analysis uses deterministic cross-directory IoU matching with class-specific source evidence. We rerun four target-removal and distractor-insertion conditions and reuse fragmentation and jitter outputs.

**Evaluation coverage.** Native inference produces 158 expression directories, of which the frozen query list admits 150. Two high-jitter queries have empty outputs and remain in evaluation as missing-target outcomes. A comparison between natural candidate sources changes several track properties simultaneously and is treated as a descriptive source comparison.

Table 12: **Corrected candidate-corruption rates and paired changes.** Intervals use three video blocks and 10,000 resamples. Baseline rates: MISS 0.3465, ASS 0.6882, DIST 0.2849, LOC 1.0000.

| Intervention | Endpoint | Dose | Rate | Paired change | 95% CI |
|---|---|---|---|---|---|
| Target removal | MISS | 0.10 | 0.4168 | +0.07027 | [+0.06435, +0.08140] |
| Target removal | MISS | 0.30 | 0.5517 | +0.20520 | [+0.18433, +0.24336] |
| ID fragmentation | ASS | 0.10 | 0.6919 | +0.00365 | [-0.00136, +0.00804] |
| ID fragmentation | ASS | 0.30 | 0.7401 | +0.05192 | [-0.00612, +0.08192] |
| Distractor insertion | DIST | 0.10 | 0.2740 | -0.01092 | [-0.01338, -0.00516] |
| Distractor insertion | DIST | 0.30 | 0.2719 | -0.01296 | [-0.03393, +0.00414] |
| Box jitter | LOC | 0.05 | 1.0000 | +0.00000 | [+0.00000, +0.00000] |
| Box jitter | LOC | 0.15 | 1.0000 | +0.00000 | [+0.00000, +0.00000] |

Table 13: **Distractor counts increase while their fraction of predictions decreases.**

| Condition | DIST events | Predictions | DIST rate |
|---|---|---|---|
| Baseline | 18,894 | 66,320 | 0.284891 |
| Low (0.10) | 20,569 | 75,077 | 0.273972 |
| High (0.30) | 23,590 | 86,749 | 0.271934 |

### E.2 RATES, COUNTS, AND PAIRED INTERVALS

**Paired inference.** Baseline and perturbed conditions share the same resampled video blocks. Each paired replicate computes

$$\Delta r_k^{(b)} = \frac{\sum_{v \in \mathcal{V}_b} N_{k,v}^{\mathrm{pert}}}{\sum_{v \in \mathcal{V}_b} D_{k,v}^{\mathrm{pert}}} - \frac{\sum_{v \in \mathcal{V}_b} N_{k,v}^{\mathrm{base}}}{\sum_{v \in \mathcal{V}_b} D_{k,v}^{\mathrm{base}}}, \tag{23}$$

where $\Delta r_k^{(b)}$ is the rate difference for error type $k$ in replicate $b$, and $\mathcal{V}_b$ is its sampled multiset of videos. For video $v$, $N_{k,v}$ and $D_{k,v}$ are the event count and denominator. Superscripts $\mathrm{pert}$ and $\mathrm{base}$ denote perturbed and baseline outputs, respectively.

Percentile intervals use 10,000 replicates with a fixed base seed. The implementation adds the sum of character code points in the concatenated condition label and player name to seed each comparison. Table 12 reports all primary rate changes. Target removal supports the positive direction at both doses. Fragmentation has monotonic point estimates with intervals crossing zero. Distractor insertion fails its positive-rate direction, and binary localization remains saturated.

**Why count and rate responses differ.** Both counts and reference populations determine rate changes:

$$\frac{N + \Delta N}{D + \Delta D} - \frac{N}{D} = \frac{D\,\Delta N - N\,\Delta D}{D(D + \Delta D)}, \tag{24}$$

where $N$ and $D$ are the baseline event count and reference population, and $\Delta N$ and $\Delta D$ are their changes under perturbation. Both $D$ and $D + \Delta D$ are positive.

For distractor insertion, the prediction denominator grows faster than the event count, producing the rate decrease in Table 13. Auxiliary continuous metrics reveal changes beyond event rates. Jitter reduces LocA from 90.872 to 82.242 and 69.856. Target removal reduces HOTA from 53.824 to 48.585 and 38.305, while fragmentation gives HOTA 53.160 and 50.032.

## F QUERY-PAIR CONSTRUCTION AND RESULTS

### F.1 REVIEW, TEST ISOLATION, AND ESTIMATOR

**Exploratory query-variation setting.** We evaluate TempRMOT, DKGTrack, and FlexHook on reviewed query pairs. Paraphrases preserve the target set, whereas single-attribute switches change it. Each output is evaluated against its own query's target ground truth. The paired estimator, hierarchical bootstrap, and joint multiple-testing correction are specified below.

Table 14: **Paraphrase results: edited minus source, averaged by source query.** Both pair types form one joint 36-test BH family, with no retained tests. HOTA is on the normalized query scale. Intervals: video then source-query resampling, 10,000 replicates, fixed seed.

| Method | Endpoint | $N$ | Estimate | 95% CI | $p$ | BH $q$ |
|---|---|---|---|---|---|---|
| TempRMOT | MISS | 24 | +0.1418 | [-0.0026, +0.2955] | 0.0538 | 0.3873 |
| TempRMOT | DIST | 24 | +0.0032 | [-0.0851, +0.0957] | 0.9385 | 0.9754 |
| TempRMOT | SEM-SUB | 24 | -0.0178 | [-0.0894, +0.0516] | 0.5975 | 0.8671 |
| TempRMOT | ASS | 24 | -0.1013 | [-0.3452, +0.1325] | 0.4062 | 0.8409 |
| TempRMOT | LOC | 24 | -0.0417 | [-0.2400, +0.1306] | 0.9017 | 0.9754 |
| TempRMOT | HOTA | 24 | -0.0935 | [-0.1942, -0.0177] | 0.0180 | 0.2160 |
| DKGTrack | MISS | 24 | +0.2152 | [+0.0479, +0.4040] | 0.0070 | 0.2160 |
| DKGTrack | DIST | 24 | -0.0694 | [-0.2400, +0.1035] | 0.4672 | 0.8409 |
| DKGTrack | SEM-SUB | 24 | -0.0628 | [-0.1660, +0.0221] | 0.1606 | 0.6423 |
| DKGTrack | ASS | 24 | -0.1490 | [-0.3461, +0.0337] | 0.1112 | 0.5003 |
| DKGTrack | LOC | 24 | -0.0833 | [-0.3077, +0.1600] | 0.6021 | 0.8671 |
| DKGTrack | HOTA | 24 | -0.1246 | [-0.2495, -0.0214] | 0.0126 | 0.2160 |
| FlexHook | MISS | 24 | +0.0012 | [-0.1048, +0.0975] | 0.8511 | 0.9575 |
| FlexHook | DIST | 24 | +0.0026 | [-0.0431, +0.0397] | 0.7871 | 0.9445 |
| FlexHook | SEM-SUB | 24 | +0.0049 | [-0.0149, +0.0292] | 0.7387 | 0.9445 |
| FlexHook | ASS | 24 | +0.0201 | [-0.0113, +0.0750] | 0.4296 | 0.8409 |
| FlexHook | LOC | 24 | +0.0417 | [+0.0000, +0.1600] | 0.9483 | 0.9754 |
| FlexHook | HOTA | 24 | -0.0014 | [-0.0426, +0.0527] | 0.8193 | 0.9515 |

**Pair construction and held-out evaluation.** Two reviewers with adjudication retain 422 of 458 constructed pairs and exclude 36. Only 47 retained pairs belong to test videos 0005, 0011, and 0013, while the remaining 375 belong to training or validation. The test set contains 25 source queries, 24 paraphrases, and 23 single-attribute target switches. No redundant-attribute deletion pair is eligible. Among 66 unique source or edited expressions, 41 reuse frozen predictions and 25 require new inference. All methods share 21,400 target detections and pass the recorded 32-coalition checks.

**Paired estimand.** For each method, we average edited-minus-source endpoint differences within source query and then across source queries:

$$\widehat{\Delta}_{e,Y} = \frac{1}{|\mathcal{Q}_e|} \sum_{q \in \mathcal{Q}_e} \frac{1}{|\mathcal{R}_e(q)|} \sum_{q' \in \mathcal{R}_e(q)} \big[Y(q') - Y(q)\big], \tag{25}$$

where $e$ is the edit type, $Y$ is the evaluated endpoint, $\mathcal{Q}_e$ contains eligible source queries $q$, and $\mathcal{R}_e(q)$ contains their reviewed rewrites $q'$. Each $Y(q)$ or $Y(q')$ uses the corresponding expression's target set, and $\widehat{\Delta}_{e,Y}$ is the mean paired effect.

The hierarchy prevents a source query with more rewrites from receiving additional weight. Bootstrap resampling draws videos and then source queries within each sampled video, using 10,000 replicates and a fixed base seed. The primary endpoints are MISS, DIST, and SEM-SUB, with ASS, LOC, and query-level HOTA as auxiliary endpoints. Event differences are rates, and HOTA differences use the normalized $[0, 1]$ query scale.

### F.2 COMPLETE MULTIPLE-TESTING FAMILY

We apply Benjamini–Hochberg correction (Benjamini & Hochberg, 1995) jointly across methods, edit types, and endpoints:

$$q_{(i)} = \min\left\{1,\ \min_{j \geq i} \frac{L}{j} p_{(j)}\right\}, \tag{26}$$

where $L = 36$ is the number of tests, $p_{(j)}$ is the $j$th smallest unadjusted value, and $q_{(i)}$ is the adjusted value at rank $i$. The inner minimum runs over all ranks satisfying $j \geq i$.

Table 14 and Table 15 report every estimate, interval, and adjusted value. Each edit type has one pair per included source query: 24 paraphrases and 23 switches, with overlap between the two source-query sets. No test survives the 0.05 FDR threshold. Some unadjusted intervals exclude zero,

Table 15: **Target-switch results: edited minus source, averaged by source query.** Both pair types form one joint 36-test BH family, with no retained tests. HOTA is on the normalized query scale. Intervals: video then source-query resampling, 10,000 replicates, fixed seed.

| Method | Endpoint | $N$ | Estimate | 95% CI | $p$ | BH $q$ |
|---|---|---|---|---|---|---|
| TempRMOT | MISS | 23 | -0.0843 | [-0.2658, +0.0789] | 0.3420 | 0.8409 |
| TempRMOT | DIST | 23 | -0.0980 | [-0.4330, +0.2348] | 0.5831 | 0.8671 |
| TempRMOT | SEM-SUB | 23 | -0.0616 | [-0.1933, +0.0446] | 0.3318 | 0.8409 |
| TempRMOT | ASS | 23 | -0.1173 | [-0.3594, +0.1497] | 0.3940 | 0.8409 |
| TempRMOT | LOC | 23 | +0.0870 | [-0.1250, +0.3182] | 0.5623 | 0.8671 |
| TempRMOT | HOTA | 23 | +0.0942 | [-0.0482, +0.2702] | 0.2508 | 0.8207 |
| DKGTrack | MISS | 23 | -0.1414 | [-0.2962, -0.0153] | 0.0248 | 0.2232 |
| DKGTrack | DIST | 23 | -0.1238 | [-0.3984, +0.1715] | 0.4100 | 0.8409 |
| DKGTrack | SEM-SUB | 23 | -0.0495 | [-0.1801, +0.0524] | 0.4040 | 0.8409 |
| DKGTrack | ASS | 23 | +0.0415 | [-0.1949, +0.2927] | 0.7653 | 0.9445 |
| DKGTrack | LOC | 23 | +0.2174 | [+0.0000, +0.4583] | 0.0956 | 0.4916 |
| DKGTrack | HOTA | 23 | +0.1190 | [-0.0108, +0.2646] | 0.0776 | 0.4656 |
| FlexHook | MISS | 23 | -0.0415 | [-0.2265, +0.1280] | 0.6429 | 0.8902 |
| FlexHook | DIST | 23 | -0.0890 | [-0.3225, +0.1401] | 0.4514 | 0.8409 |
| FlexHook | SEM-SUB | 23 | -0.0620 | [-0.1589, +0.0295] | 0.1942 | 0.6991 |
| FlexHook | ASS | 23 | -0.0355 | [-0.2560, +0.2450] | 0.7143 | 0.9445 |
| FlexHook | LOC | 23 | +0.0000 | [-0.2174, +0.1905] | 1.0000 | 1.0000 |
| FlexHook | HOTA | 23 | +0.0466 | [-0.1182, +0.2139] | 0.5811 | 0.8671 |

but the family-level result remains unchanged. We do not test equivalence. **Meaning-preserving paraphrases.** Table 14 compares paraphrases with source queries under the same target set. No endpoint survives joint correction, which does not establish wording equivalence.

**Single-attribute target switches.** Table 15 evaluates edits that change the requested objects. Each output is scored against its own target set, so these comparisons measure responses to changed requests rather than invariance to wording.

# G REFER-KITTI-V2 REUSE

## G.1 REGISTRY, CHECKPOINTS, AND PROFILES

**Registry and outputs.** The four test videos contribute 197, 228, 207, and 230 queries, for 862 queries and 491,222 target detections. Eighty-eight query-frame disagreements are resolved against the official TrackEval GT template. The constant-zero category field supports target/distractor membership but not category-specific interpretation. TempRMOT produces 55 valid empty-query outputs, whereas FlexHook produces none. Both use the released v2 checkpoints and evaluate all 32 repair coalitions plus an independently constructed ideal.

**Diagnostic profiles.** Table 16 reports the HOTA budgets, and Table 17 gives all pairwise interactions. Both methods retain the largest MISS allocation and the two prespecified positive interaction signs. Frozen TempRMOT obtains a HOTA score of 34.345, versus the public score of 35.040. Resolving this 0.695-point gap requires prediction-level public provenance. FlexHook reproduces 42.526. Dataset-specific checkpoints make this a test of protocol reuse, rather than zero-shot transfer of one RMOT method. **Cross-dataset budgets.** As shown in Table 16, missing targets account for the largest allocation in both methods, while their secondary budgets differ. The zero residuals indicate complete recovery within the defined repair family.

**Cross-dataset interactions.** Table 17 gives all pairwise indices underlying the comparison in Fig. 6(b). The prespecified pairs retain positive indices, placing the selected complementary gains in the context of the complete interaction profile.

Table 16: **Refer-KITTI-v2 HOTA budgets under dataset-specific checkpoints.**

| Method | Raw | MISS | DIST | BKG-DUP | ASS | LOC | RES |
|---|---|---|---|---|---|---|---|
| TempRMOT | 34.345 | 47.606 | 9.235 | 1.412 | 3.965 | 3.438 | 0.000 |
| FlexHook | 42.526 | 23.472 | 17.489 | 4.206 | 4.924 | 7.383 | 0.000 |

Table 17: **All Refer-KITTI-v2 HOTA interaction indices.**

| Pair | TempRMOT | FlexHook |
|---|---|---|
| MISS–DIST | **+6.485** | **+5.996** |
| MISS–BKG-DUP | +0.940 | +1.356 |
| MISS–ASS | +3.048 | +2.071 |
| MISS–LOC | -0.123 | +0.560 |
| DIST–BKG-DUP | +0.549 | +2.908 |
| DIST–ASS | -0.139 | +0.364 |
| DIST–LOC | +0.640 | +2.408 |
| BKG-DUP–ASS | -0.079 | -0.044 |
| BKG-DUP–LOC | +0.113 | +0.614 |
| ASS–LOC | +0.875 | +1.626 |

### G.2 EXPLORATORY DIFFICULTY ASSOCIATIONS

**Model specification.** We model 1,724 method-query rows with standardized query and scene features, method-specific slopes, and penalized video and query-template intercepts:

$$\widehat{Y}_{mq} = \beta_0 + \beta_m h_m + \mathbf{z}_q^\top \boldsymbol{\beta} + h_m \mathbf{z}_q^\top \boldsymbol{\gamma} + u_{v(q)} + w_{a(q)}, \tag{27}$$

where $\widehat{Y}_{mq}$ is the fitted endpoint for method $m$ and query $q$, and $h_m$ is one for TempRMOT and zero for FlexHook. The vector $\mathbf{z}_q$ contains nine standardized predictors. The coefficients $\beta_0$, $\beta_m$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ denote the intercept, method offset, FlexHook slopes, and TempRMOT slope differences, respectively. The terms $u_{v(q)}$ and $w_{a(q)}$ are intercepts for the query's video $v(q)$ and template $a(q)$.

We fit the coefficients by minimizing

$$\sum_{m,q} \left(Y_{mq} - \widehat{Y}_{mq}\right)^2 + \lambda \left( \sum_v u_v^2 + \sum_a w_a^2 \right), \qquad \lambda = 1, \tag{28}$$

where $Y_{mq}$ is the observed endpoint, $\widehat{Y}_{mq}$ its fitted value, and $\lambda$ controls the squared penalties on video effects $u_v$ and template effects $w_a$. The first sum covers method-query rows, and the penalty sums cover all videos $v$ and templates $a$.

Predictors describe attribute count, spatial and motion language, target count and area, distractor count, crowding, displacement, and CLIP similarity to alternative expressions. With constant-zero category labels, the same-label distractor-count feature reduces to annotated distractor count in valid target frames. It cannot distinguish cars from pedestrians.

**Retained associations.** Inference uses 1,000 video-block bootstrap replicates and one 72-term BH family. Four terms remain at $q = 0.036$. Target count is associated with DIST rate ($-0.0950$, 95% CI $[-0.2582, -0.0625]$), as is displacement for FlexHook ($+0.0940$, $[+0.0185, +0.2135]$). The TempRMOT-by-displacement interaction is negative ($-0.0765$, $[-0.2117, -0.0464]$), while crowding is positively associated with SEM-SUB rate ($+0.0360$, $[+0.0145, +0.1452]$). Equation (27) gives TempRMOT's total slope as $\beta_j + \gamma_j$.

**Stability and scope.** These associations are exploratory, with no independent predictive validation and only four video clusters. Neither HOTA nor MISS-rate predictors pass the FDR threshold, so the model does not explain the dominant missing-target allocation.

Table 18: **Leave-one-video-out coefficients for the four FDR-retained terms.** Column labels identify the omitted video. Sign stability is a within-sample diagnostic.

| Outcome | Term | 0005 | 0011 | 0013 | 0019 |
|---|---|---|---|---|---|
| DIST_rate | Target count | -0.1136 | -0.1413 | -0.0860 | -0.0724 |
| DIST_rate | Normalized displacement | +0.0549 | +0.1974 | +0.0733 | +0.1063 |
| DIST_rate | TempRMOT × Normalized displacement | -0.0880 | -0.1148 | -0.0555 | -0.0674 |
| SEM_SUB_rate | Crowding IoU proxy | +0.0454 | +0.0445 | +0.0327 | +0.0404 |

Table 19: **Complete difficulty coefficients for HOTA.** Predictors are standardized. TempRMOT interactions are slope differences relative to FlexHook. Intercepts and method offsets are excluded from the joint 72-term BH family.

| Term | Coefficient | 95% CI | $p$ | BH $q$ |
|---|---|---|---|---|
| Intercept | +38.0025 | [+29.6373, +58.5727] | 0.0020 | – |
| TempRMOT offset | -10.6789 | [-25.5413, -8.6821] | 0.0020 | – |
| Attribute count | -2.3047 | [-4.2473, +3.0875] | 0.5734 | 0.8631 |
| Spatial relation | +4.1484 | [-5.1750, +6.4302] | 0.1758 | 0.4980 |
| Motion relation | +0.3637 | [-3.4536, +0.6387] | 0.6334 | 0.8631 |
| Target count | +6.1563 | [+4.1127, +24.3265] | 0.0080 | 0.0575 |
| Same-label distractor count | -3.0072 | [-14.7039, +4.9765] | 0.7453 | 0.8696 |
| Target area ratio | -0.8839 | [-5.2997, +4.6410] | 0.9091 | 0.9091 |
| Normalized displacement | -8.8410 | [-18.6888, -2.4161] | 0.0160 | 0.1046 |
| Crowding IoU proxy | -0.5813 | [-9.0605, +4.4575] | 0.5435 | 0.8631 |
| CLIP alternative similarity | +2.0561 | [-0.8126, +4.0472] | 0.1638 | 0.4975 |
| TempRMOT × Attribute count | -1.7573 | [-6.8835, +0.5355] | 0.0819 | 0.3686 |
| TempRMOT × Spatial relation | +1.4359 | [-2.7410, +5.9790] | 0.3836 | 0.7792 |
| TempRMOT × Motion relation | +2.3342 | [+0.7349, +4.8354] | 0.0080 | 0.0575 |
| TempRMOT × Target count | -0.4548 | [-18.2502, +2.2951] | 0.7632 | 0.8696 |
| TempRMOT × Same-label distractor count | -1.0918 | [-9.0093, +2.1331] | 0.5155 | 0.8435 |
| TempRMOT × Target area ratio | +3.9221 | [-5.0000, +5.8075] | 0.1439 | 0.4932 |
| TempRMOT × Normalized displacement | +3.7882 | [-0.4897, +5.0860] | 0.2817 | 0.7244 |
| TempRMOT × Crowding IoU proxy | +0.7080 | [-8.9597, +7.5461] | 0.6054 | 0.8631 |
| TempRMOT × CLIP alternative similarity | -1.5328 | [-3.7172, +3.9259] | 0.6254 | 0.8631 |

### G.3 Sensitivity to Individual Videos

Table 18 checks the retained associations after omitting each video in turn. Their signs are stable across these fits, although the analysis still covers the same small set of scenes and does not provide independent predictive validation.

### G.4 Difficulty Results by Endpoint

**Aggregate tracking quality.** Table 19 reports the HOTA regression. Some unadjusted intervals exclude zero, but no predictor survives the common FDR correction. These fits do not establish a general explanation of tracking difficulty.

**Missing-target occurrence.** As shown in Table 20, no predictor of missing-target rate survives the same correction. This occurrence regression does not explain the large missing-target allocation, which measures a different diagnostic quantity.

**Distractor occurrence.** Table 21 identifies the retained target-count, displacement, and method-by-displacement associations. The interaction coefficient describes a difference between methods' slopes, while each method's total slope includes its main and interaction terms.

**Missing targets and distractors in the same frame.** Table 22 reports the association between crowding and SEM-SUB rate. This endpoint records co-occurrence and does not identify a substituted object pair or an internal semantic failure.

Table 20: **Complete difficulty coefficients for MISS_rate.** Predictors are standardized. TempRMOT interactions are slope differences relative to FlexHook. Intercepts and method offsets are excluded from the joint 72-term BH family.

| Term | Coefficient | 95% CI | $p$ | BH $q$ |
|---|---|---|---|---|
| Intercept | +0.4308 | [+0.2689, +0.6299] | 0.0100 | – |
| TempRMOT offset | +0.2453 | [+0.1727, +0.3018] | 0.0020 | – |
| Attribute count | +0.0541 | [-0.0522, +0.0963] | 0.4416 | 0.8000 |
| Spatial relation | -0.0640 | [-0.1263, +0.1081] | 0.4036 | 0.7854 |
| Motion relation | -0.0125 | [-0.0232, +0.0521] | 0.8511 | 0.8755 |
| Target count | -0.0432 | [-0.1087, -0.0144] | 0.0400 | 0.2213 |
| Same-label distractor count | +0.0770 | [-0.0099, +0.2700] | 0.1339 | 0.4891 |
| Target area ratio | +0.0125 | [-0.0895, +0.0902] | 0.7772 | 0.8696 |
| Normalized displacement | +0.1009 | [-0.0499, +0.1804] | 0.1658 | 0.4975 |
| Crowding IoU proxy | -0.0354 | [-0.1375, +0.1148] | 0.7652 | 0.8696 |
| CLIP alternative similarity | -0.0331 | [-0.0662, +0.0346] | 0.4555 | 0.8000 |
| TempRMOT × Attribute count | +0.0034 | [-0.0303, +0.1195] | 0.6194 | 0.8631 |
| TempRMOT × Spatial relation | -0.0011 | [-0.0908, +0.0621] | 0.8432 | 0.8755 |
| TempRMOT × Motion relation | -0.0278 | [-0.0527, -0.0129] | 0.0519 | 0.2672 |
| TempRMOT × Target count | -0.0087 | [-0.0439, +0.0926] | 0.8092 | 0.8696 |
| TempRMOT × Same-label distractor count | -0.0657 | [-0.1272, +0.1163] | 0.4675 | 0.8015 |
| TempRMOT × Target area ratio | -0.0526 | [-0.0857, +0.0877] | 0.1518 | 0.4970 |
| TempRMOT × Normalized displacement | -0.0491 | [-0.0937, +0.0629] | 0.4975 | 0.8330 |
| TempRMOT × Crowding IoU proxy | +0.0687 | [-0.0924, +0.1850] | 0.3776 | 0.7792 |
| TempRMOT × CLIP alternative similarity | +0.0230 | [-0.0308, +0.0498] | 0.6474 | 0.8631 |

Table 21: **Complete difficulty coefficients for DIST_rate.** Predictors are standardized. TempRMOT interactions are slope differences relative to FlexHook. Intercepts and method offsets are excluded from the joint 72-term BH family.

| Term | Coefficient | 95% CI | $p$ | BH $q$ |
|---|---|---|---|---|
| Intercept | +0.4382 | [+0.3217, +0.5747] | 0.0080 | – |
| TempRMOT offset | +0.0091 | [-0.0481, +0.0740] | 0.7772 | – |
| Attribute count | -0.0426 | [-0.1275, +0.0495] | 0.4535 | 0.8000 |
| Spatial relation | -0.0565 | [-0.0890, +0.0363] | 0.0999 | 0.4231 |
| Motion relation | +0.0210 | [-0.0372, +0.1365] | 0.5814 | 0.8631 |
| Target count | -0.0950 | [-0.2582, -0.0625] | 0.0020 | 0.0360 |
| Same-label distractor count | +0.0872 | [-0.1349, +0.1685] | 0.3536 | 0.7792 |
| Target area ratio | +0.0647 | [+0.0238, +0.1186] | 0.0080 | 0.0575 |
| Normalized displacement | +0.0940 | [+0.0185, +0.2135] | 0.0020 | 0.0360 |
| Crowding IoU proxy | +0.0189 | [-0.0196, +0.3599] | 0.3337 | 0.7792 |
| CLIP alternative similarity | +0.0392 | [-0.0173, +0.0586] | 0.3896 | 0.7792 |
| TempRMOT × Attribute count | +0.0274 | [-0.0558, +0.0788] | 0.6314 | 0.8631 |
| TempRMOT × Spatial relation | +0.0005 | [-0.0265, +0.0570] | 0.7532 | 0.8696 |
| TempRMOT × Motion relation | -0.0129 | [-0.0464, +0.0529] | 0.8931 | 0.9057 |
| TempRMOT × Target count | -0.0003 | [-0.0294, +0.1567] | 0.8332 | 0.8755 |
| TempRMOT × Same-label distractor count | -0.0456 | [-0.0618, +0.0748] | 0.6394 | 0.8631 |
| TempRMOT × Target area ratio | -0.0347 | [-0.0739, -0.0043] | 0.0080 | 0.0575 |
| TempRMOT × Normalized displacement | -0.0765 | [-0.2117, -0.0464] | 0.0020 | 0.0360 |
| TempRMOT × Crowding IoU proxy | -0.0215 | [-0.1966, +0.0085] | 0.1798 | 0.4980 |
| TempRMOT × CLIP alternative similarity | +0.0063 | [-0.0090, +0.0392] | 0.3796 | 0.7792 |

Table 22: **Complete difficulty coefficients for SEM_SUB_rate.** Predictors are standardized. TempRMOT interactions are slope differences relative to FlexHook. Intercepts and method offsets are excluded from the joint 72-term BH family.

| Term | Coefficient | 95% CI | $p$ | BH $q$ |
|---|---|---|---|---|
| Intercept | +0.1849 | [+0.0946, +0.2431] | 0.0020 | – |
| TempRMOT offset | -0.0169 | [-0.0652, +0.1025] | 0.3397 | – |
| Attribute count | -0.0193 | [-0.0665, +0.0284] | 0.3716 | 0.7792 |
| Spatial relation | -0.0383 | [-0.0514, -0.0057] | 0.0200 | 0.1199 |
| Motion relation | +0.0053 | [-0.0583, +0.0733] | 0.7512 | 0.8696 |
| Target count | +0.0514 | [-0.1066, +0.0732] | 0.1978 | 0.5275 |
| Same-label distractor count | +0.0352 | [-0.0612, +0.0971] | 0.4316 | 0.8000 |
| Target area ratio | -0.0044 | [-0.0345, +0.0377] | 0.8052 | 0.8696 |
| Normalized displacement | +0.0057 | [-0.0392, +0.1153] | 0.6713 | 0.8631 |
| Crowding IoU proxy | +0.0360 | [+0.0145, +0.1452] | 0.0020 | 0.0360 |
| CLIP alternative similarity | +0.0019 | [-0.0123, +0.0098] | 0.8072 | 0.8696 |
| TempRMOT × Attribute count | -0.0248 | [-0.0652, +0.0013] | 0.0699 | 0.3357 |
| TempRMOT × Spatial relation | +0.0341 | [+0.0244, +0.0561] | 0.0080 | 0.0575 |
| TempRMOT × Motion relation | +0.0325 | [-0.0085, +0.0773] | 0.1359 | 0.4891 |
| TempRMOT × Target count | -0.0133 | [-0.0820, +0.1348] | 0.6893 | 0.8696 |
| TempRMOT × Same-label distractor count | +0.0136 | [-0.0818, +0.1538] | 0.3576 | 0.7792 |
| TempRMOT × Target area ratio | -0.0012 | [-0.0493, +0.0254] | 0.7792 | 0.8696 |
| TempRMOT × Normalized displacement | -0.0247 | [-0.2011, +0.0033] | 0.1119 | 0.4476 |
| TempRMOT × Crowding IoU proxy | -0.0362 | [-0.1372, -0.0164] | 0.0080 | 0.0575 |
| TempRMOT × CLIP alternative similarity | +0.0040 | [-0.0103, +0.0312] | 0.6713 | 0.8631 |