

# AerialMind: Towards Referring Multi-Object Tracking in UAV Scenarios

Chenglizhao Chen<sup>1,2</sup>, Shaofeng Liang<sup>1,2</sup>, Runwei Guan<sup>3 \*</sup>, Xiaolou Sun<sup>4</sup>, Haocheng Zhao<sup>5</sup>,  
Haiyun Jiang<sup>6</sup>, Tao Huang<sup>7</sup>, Henghui Ding<sup>8</sup>, Qing-Long Han<sup>9</sup>

<sup>1</sup>Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China)

<sup>2</sup>Shandong Key Laboratory of Intelligent Oil & Gas Industrial Software

<sup>3</sup>Thrust of Artificial Intelligence, The Hong Kong University of Science and Technology (Guangzhou)

<sup>4</sup>Purple Mountain Laboratories

<sup>5</sup>School of Advanced Technology, Xi'an Jiaotong-Liverpool University

<sup>6</sup>School of Automation and Intelligent Sensing, Shanghai Jiao Tong University

<sup>7</sup>College of Science and Engineering, James Cook University

<sup>8</sup>Institute of Big Data, College of Computer Science and Artificial Intelligence, Fudan University

<sup>9</sup>School of Engineering, Swinburne University of Technology, Melbourne

## Abstract

Referring Multi-Object Tracking (RMOT) aims to achieve precise object detection and tracking through natural language instructions, representing a fundamental capability for intelligent robotic systems. However, current RMOT research remains mostly confined to ground-level scenarios, which constrains their ability to capture broad-scale scene contexts and perform comprehensive tracking and path planning. In contrast, Unmanned Aerial Vehicles (UAVs) leverage their expansive aerial perspectives and superior maneuverability to enable wide-area surveillance. Moreover, UAVs have emerged as critical platforms for Embodied Intelligence, which has given rise to an unprecedented demand for intelligent aerial systems capable of natural language interaction. To this end, we introduce AerialMind, the first large-scale RMOT benchmark in UAV scenarios, which aims to bridge this research gap. To facilitate its construction, we develop an innovative semi-automated collaborative agent-based labeling assistant (COALA) framework that significantly reduces labor costs while maintaining annotation quality. Furthermore, we propose HawkEyeTrack (HETrack), a novel method that collaboratively enhances vision-language representation learning and improves the perception of UAV scenarios. Comprehensive experiments validated the challenging nature of our dataset and the effectiveness of our method.

Datasets — <https://github.com/shawnliang420/AerialMind>

## Introduction

Referring Multi-Object Tracking (RMOT) (Wu et al. 2023; Zhang et al. 2024) aims to achieve precise detection and tracking of specified targets in video sequences through language instructions. It realizes a fundamental paradigm shift from passive perception to active understanding. Although significant progress (Du et al. 2024; Chen et al. 2025a; Ma et al. 2024; Wu et al. 2023; Chen et al. 2024a)



Figure 1: Overview of the challenges in AerialMind dataset.

has been achieved, it is almost entirely confined to ground-level scenarios. It constrains their ability to capture broad-scale scene contexts and perform comprehensive tracking and path planning. In contrast, Unmanned Aerial Vehicles (UAVs) leverage expansive aerial perspectives and superior maneuverability to enable wide-area surveillance capabilities unattainable by ground-based systems. As critical platforms for Embodied AI (Wang et al. 2025), UAVs drive unprecedented demand for intelligent aerial systems with natural language interaction capabilities. However, current RMOT research lacks sufficient exploration of challenging aerial scenarios, resulting in limited real-world applicability and hindering the realization of truly aerial intelligence.

To this end, we construct the first large-scale referring multi-object tracking dataset **AerialMind** for UAV scenarios. The dataset is extended based on VisDrone (Du et al. 2019) and UAVDT (Du et al. 2018), covering multiple flight altitudes, environmental conditions, and target categories. As shown in Figure 1, AerialMind brings unprecedented challenges: **①** Drastic Appearance Difference: Changes in flight altitude and viewpoints cause dramatic differences in object appearance; **②** Complex Spatial Relationships: Object relationships under aerial view perspectives are more

\*Corresponding author: runwayrwguan@hkust-gz.edu.cn  
Copyright © 2026, Association for the Advancement of Artificial Intelligence ([www.aaai.org](http://www.aaai.org)). All rights reserved.

intricate; ③ Dynamic Scene Variations: The high maneuverability of UAVs brings continuously changing scenes and illumination conditions; ④ Various Referring Expressions: Spatial, motion states, and object descriptions in UAV scenarios exhibit richer semantic complexity. To facilitate profound and quantitative analysis, we also pioneer frame-by-frame attribute annotations in the RMOT field.

To efficiently construct AerialMind, we develop a novel semi-automated annotation framework, namely **CO**llaborative **A**gent-based **L**abeling **A**ssistant (COALA). It aims to reduce annotation costs through intelligent processes while effectively avoiding subjective biases in manual annotation. Specifically, COALA adopts a multi-stage annotation mechanism: First, it utilizes large language models (LLMs) to intelligently parse UAV scenarios; Then, the system automatically records targeted objects by annotators simply click and define the temporal boundaries of referring events, and associates corresponding description items; Subsequently, it performs cross-modal logical reasoning on static frames and trajectory data to validate annotation quality. Finally, it leverages the generative capabilities of LLMs to expand and generate more semantically rich expressions.

Furthermore, we propose a novel method called HawkEyeTrack (HETrack). It innovatively introduces the Co-evolutionary Fusion Encoder (CFE) that enables a co-evolutionary refinement of vision and language representations and incorporates a targeted Scale Adaptive Contextual Refinement (SACR) module to significantly enhance the perception of UAV scenarios. Comprehensive experiments on AerialMind validate the challenging nature of the benchmark and demonstrate the effectiveness of HETrack.

In summary, our contributions are listed as follows:

1. **AerialMind benchmark dataset:** We construct the first large-scale referring multi-object tracking benchmark dataset for Unmanned Aerial Vehicle (UAV) scenarios. It introduces new challenges for RMOT research.
2. **COALA annotation framework:** An innovative semi-automated annotation framework that adopts multi-stage agent collaborative mechanisms, significantly reducing manual costs while ensuring high-quality annotations.
3. **HETrack method:** It integrates the co-evolutionary refinement of vision and language representations and scale adaptive contextual refinement, achieving excellent performance on our AerialMind dataset.

## Related Works

### Referring Understanding Datasets

Referring to understanding tasks (Ding et al. 2022a, 2023, 2025a,b; Guan et al. 2024, 2025c,b), which aim to localize specific regions in images or videos through natural language expressions. Early dataset construction work mainly focused on static image scenarios, such as the RefCOCO (Yu et al. 2016) series datasets. Subsequently, researchers gradually extended referring understanding to temporal video domains, successively proposing video referring segmentation datasets such as Refer-DAVIS<sub>17</sub> (Khoreva, Rohrbach, and Schiele 2019) and Refer-Youtube-VOS (Seo, Lee, and Han

2020). Wu et al. (Wu et al. 2023) first proposed the referring multi-object tracking task. Researchers further extended this work, proposing larger-scale Refer-KITTI-V2 (Zhang et al. 2024), Refer-BDD (Chen et al. 2025a) and ReaMOT (Chen et al. 2025b) datasets. They mainly focus on specific ground perspectives, lacking sufficient consideration for the unique challenges of aerial platforms such as UAVs. Recently, Researchers (Sun et al. 2025; Liu et al. 2025) constructed the UAV referring expression detection datasets, validating the feasibility of referring understanding from aerial perspectives. However, they concentrate on single-frame detection tasks, lacking in-depth exploration that requires long-term temporal modeling and complex language understanding.

### Referring Understanding Methods

Early referring understanding methods (Khoreva, Rohrbach, and Schiele 2019; Luo and Shakhnarovich 2017) mostly adopted two-stage strategies (Zhou et al. 2022), which rely heavily on candidate region quality and have low computational efficiency. Currently, end-to-end methods (Liang et al. 2023; Liao et al. 2020) have gradually become mainstream. These methods achieve visual-language fusion through designing sophisticated mechanisms (Luo et al. 2020; Sun et al. 2020; Ding et al. 2022b; Hui et al. 2021; Wu et al. 2022). For referring to multi-object tracking, TransRMOT (Wu et al. 2023) first proposed an end-to-end solution based on the Transformer. TempRMOT (Zhang et al. 2024) further introduced temporal enhancement modules, improving the temporal consistency of tracking. Although these methods (Du et al. 2024; Chen et al. 2025a) have achieved significant progress in ground scenarios, they still show inadequate adaptability when facing unique UAV challenges.

### Benchmark

We construct the first large-scale referring multi-object tracking benchmark **AerialMind** for unmanned aerial vehicle (UAV) scenarios. We demonstrate the core challenges presented by AerialMind in Figure 1 and provide detailed data statistical analysis in Figure 2.

### Dataset Features and Statistics

As shown in Table 1, AerialMind contains 93 video sequences, totaling 24.6K referring expressions, associated with 293.1K object instances and up to 46.14M bounding box annotations. In comparison, even the larger-scale Refer-KITTI-V2 has only 9.8K expressions, less than half of AerialMind. More importantly, AerialMind systematically covers cross-domain scenarios and complex referring expressions (including 752 no-target expressions and 458 reasoning expressions) and fine-grained attribute annotations, greatly enhancing the comprehensive challenge of the task. The word clouds and semantic concepts are as shown in Figure 2-a & e, demonstrating the rich linguistic diversity and semantic breadth of our dataset. The temporal ratio distribution of referring expressions in videos (Figure 2-c) is broad and balanced, meaning referring events may occur or end at any point in the video. The representative frame count

Dataset	Source	Videos	Dom.	Reas.	Attr.	Expressions	Words	Instance / Expression	Instances	Bbox Anno.
Refer-KITTI	CVPR2023	18	X	X	X	818	49	10.7	8.8K	0.36M
Refer-Dance	CVPR2024	65	X	X	X	1.9K	25	0.34	650	0.55M
Refer-KITTI-V2	arXiv2024	21	X	X	X	9.8K	617	6.7	65.4K	3.06M
Refer-UE-City	arXiv2024	12	X	X	X	714	—	10.3	—	0.55M
Refer-BDD	IEEE TIM2025	50	X	X	X	4.6K	225	15.3	70.4K	1.50M
CRTrack	AAAI2025	41	✓	X	X	344	43	—	—	0.79M
LaMOT*	IEEE ICRA2025	62	X	X	X	145	9	54.6	7.9K	1.2M
AerialMind	Ours	93	✓	✓	✓	24.6K	1.2K	11.9	293.1K	46.14M

Table 1: Comparison of referring multi-object tracking datasets. The Dom. represents cross-domain scenarios, Reas. denotes complex reasoning expressions, and Attr. indicates attribute annotation. LaMOT\* (Li et al. 2025a) represents the UAV subset.

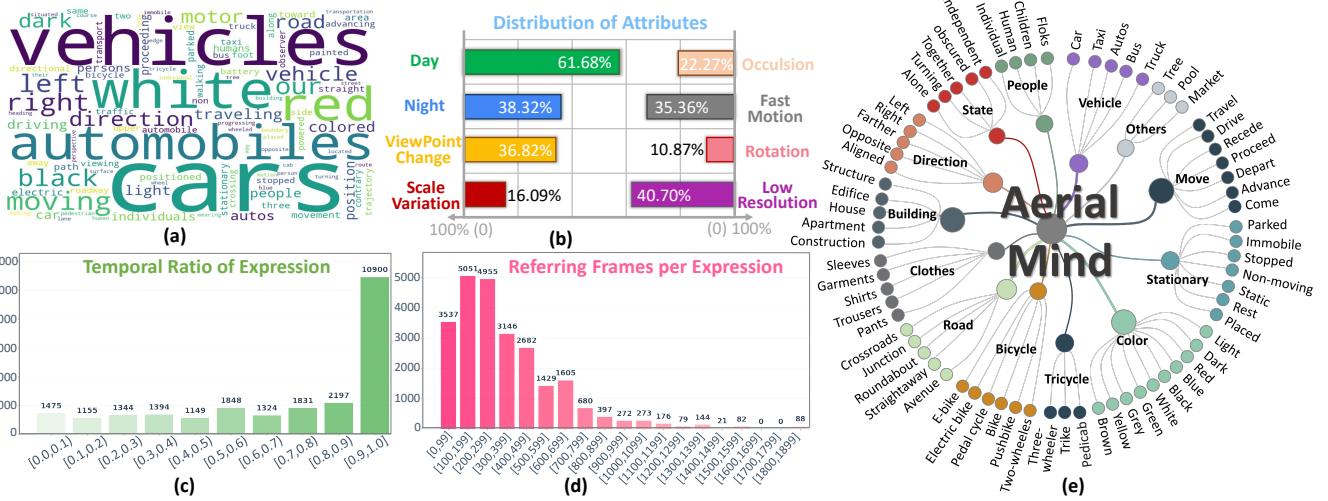


Figure 2: Overview of the AerialMind dataset statistics. It shows the distribution and diversity of (a) vocabulary, (b) challenging attributes, (c-d) temporal characteristics, and (e) semantic concepts.

distribution (see Figure 2-d) exhibits obvious long-tail characteristics, with many long-term referring events spanning hundreds of frames. This presents significant challenges for models’ temporal event localization abilities.

To promote deeper and more refined diagnostic analysis of model performance, we introduce a novel attribute-based evaluation, which is the first exploration in the RMOT field. We frame-by-frame annotate eight challenge attributes in the test set: illumination conditions (day/night), viewpoint change, scale variation, occlusion, fast motion, camera rotation, and low resolution, as shown in Figure 2-b. We introduced new metrics, namely HOTA<sub>S</sub> and HOTA<sub>M</sub>, to evaluate the model’s capability in addressing scene-induced and motion-induced challenges, respectively.

### Collaborative Agent-based Labeling Assistant

Traditional annotation pipelines for referring expressions are labor-intensive and time-costly. Consequently, we introduce the **CO**llaborative **A**gent-based **L**abeling **A**sistant (**COALA**), a novel semi-automated framework designed to augment the traditional annotation pipeline through the LLM agent. It explores a sustainable and scalable next-generation visual annotation paradigm that achieves a bal-

ance between cost, efficiency, quality, and diversity.

**1. Scene Understanding and Prompt Generation:** To alleviate the high cognitive load and time costs caused by tedious manual video review, we first utilize the few-shot visual question answering (VQA) capabilities of large language models (LLMs) to guide the annotation process. For each video, we design a Scene Understanding Prompt Agent (SUP-Agent). This agent takes key frames of the video as input and automatically generates high-level semantic digests of the scene. It covers descriptions of key objects, attributes, and spatial layouts and includes a series of templated scene prompts, providing a structured starting point for subsequent annotation tasks, as shown in Figure 3-Stage 1.

**2. Semi-automated Object Labeling:** Figure 3-Stage 2 shows that we introduce a human-machine collaborative workflow, whose core is the Semi-automated Object Labeling Agent (SOL-Agent). The human annotator first reads the digest from Stage 1 to gain a comprehensive understanding of the video’s context. Subsequently, the annotator actively selects a language description from the templated scene prompts. With the chosen text as guidance, annotators only need to perform two clicks on a target instance to define the complete temporal interval (i.e., the start and

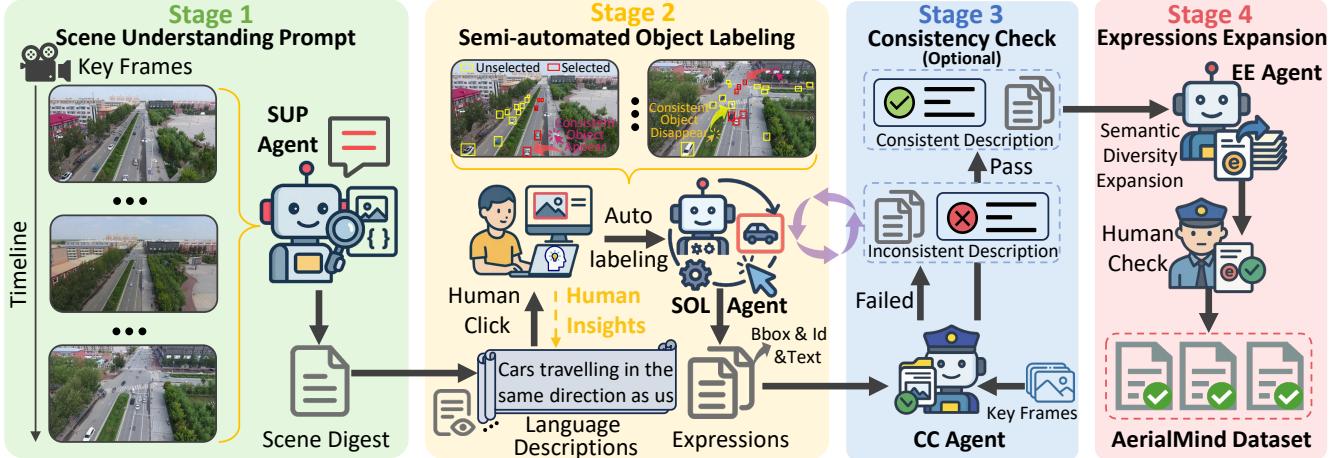


Figure 3: Overview of the four-stage annotation process in the COALA framework. This framework efficiently constructs the AerialMind dataset through multi-agent collaboration and human-computer interaction.

end points) where it matches the description. The SOL-Agent then tracks and associates its corresponding bounding box trajectory frame by frame based on the existing detection boxes in the video. This “click-to-define” interaction paradigm liberates annotators from tedious frame-by-frame operations, allowing them to focus on high-level semantic judgment and temporal boundary definition, greatly improving annotation efficiency and consistency. More importantly, when human experts identify more complex or subtle interactions that are not covered by preset scene prompts during the annotation process. They can directly create new, more precise linguistic descriptions instantly, ensuring comprehensive coverage of complex real-world scenarios.

**3. Consistency Check:** To ensure the highest quality of annotations and lay the foundation for future fully automated processes, we introduce an optional but crucial Consistency Check Agent (CC-Agent), as illustrated in Figure 3-Stage 3. Its core innovation is to perform cross-modal spatio-temporal logical reasoning by analyzing a comprehensive data package based on LLM. Specifically, the CC-Agent validates the matching degree between visual features, language descriptions, and the motion patterns inferred from trajectory data (such as velocity and directional changes). Annotations that fail to pass validation will be returned for correction. This stage is designated as optional, primarily to balance its significant cost against the already high fidelity of the preceding annotations. However, it serves as the foundation for a fully automated annotation in the feature.

**4. Expression Expansion:** In the final stage, we design an Expression Expansion Agent (EE-Agent) that plays the role of a “linguist” (see Figure 3-Stage 4). This agent takes validated expressions as “semantic seeds” and is prompted to generate multiple new expressions that differ in syntax and vocabulary but are semantically equivalent. This step greatly enriches the linguistic diversity of the dataset. Finally, all machine-generated expressions enter a final human verification process to thoroughly filter out any potential errors or “hallucinations” introduced by LLMs.

## Method

In this work, we propose a novel framework named HawkEyeTrack (HETrack) for robust referring tracking. We introduce two key innovations: a Co-evolutionary Fusion Encoder that enables collaborative refinement of vision and language representations, and a Scale Adaptive Contextual Refinement module to significantly enhance the perception of UAV scenarios, as shown in Figure 4.

### Co-evolutionary Fusion Encoder

In language-guided visual perception tasks, achieving efficient Cross-Modal Representation Alignment (Chen et al. 2024b, 2025d,c) is the core challenge. Existing methods mostly follow the early fusion and late fusion paradigms. Early fusion attempts to forcibly align highly abstract text with unstructured, noisy visual features at the beginning of visual encoding. It faces a huge modality gap and may cause the language signal to be progressively diluted in the subsequent encoding. Conversely, although late fusion structures the structural visual features, it makes this process a blind exploration without language navigation, resulting in the final fusion to an inefficient “post-hoc correction”. These become increasingly prominent when facing various descriptions in AerialMind that are full of complex spatial relationships. To this end, we propose a novel Co-evolutionary Fusion Encoder (CFE). Our key insight is: the structuring process of visual features and the guiding process of language information should not be independent stages, but rather a deeply intertwined and mutually reinforcing unified body.

Specifically, given an image frame, a visual backbone network extracts a multi-scale feature pyramid, denoted as  $\mathbf{F}_V = \{\mathbf{V}^{(l)}\}_{l=1}^{L_s}$ , where  $\mathbf{V}^{(l)} \in \mathbb{R}^{H_l \times W_l \times C_l}$  represents the feature map at the  $l$ -th level. Concurrently, the input language expression is encoded via a text encoder into two granularities: word-level features  $\mathbf{T}_w \in \mathbb{R}^{L \times C}$  and a sentence-level global feature  $\mathbf{T}_s \in \mathbb{R}^{1 \times C}$ . The CFE is constructed by stacking  $N_e$  blocks. Each block comprises

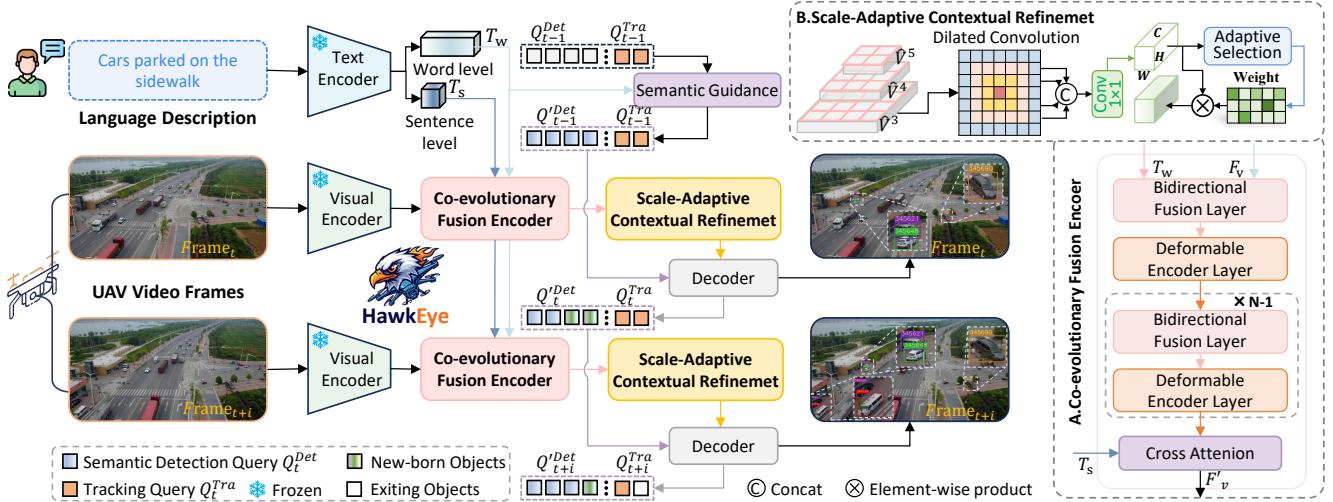


Figure 4: Overview of the HawkEyeTrack. Our key innovations include the Co-evolutionary Fusion Encoder for synergistic vision-language alignment and Scale-Adaptive Contextual Refinement for enhancing the perception of UAV scenarios.

a Bidirectional Fusion Layer (BFL) and a Deformable Encoder Layer (DEL). The bidirectional nature of this fusion implies that visual features provide concrete anchors for linguistic concepts, while linguistic concepts offer targeted guidance for the filtering and enhancement of visual features. Then, the fused features  $\mathbf{F}_V^{(i)}$  are immediately processed by a DEL for efficient intra-modal spatial relationship modeling. After  $N_e$  iterations, we obtain a final visual representation,  $\mathbf{F}_{enc} = \mathbf{F}_V^{(N_e)}$ . To imbue the model with a holistic grasp of the overall referring intent, we leverage the global sentence-level feature  $\mathbf{T}_s$  to perform a final modulation on the co-evolved visual features  $\hat{\mathbf{F}}_V$ . Formally:

$$\begin{aligned} \mathbf{F}_V^{(i)}; \mathbf{T}_w^{(i)} &= \text{BFL}^i(\mathbf{F}_V^{(i)}, \mathbf{T}_w^{(i)}; \theta_i) \\ &= \mathbf{F}_V^{(i)} + \underbrace{\Delta \mathbf{F}_V^{(i)}; \mathbf{T}_w^{(i)} + \Delta \mathbf{T}_w^{(i)}}_{\text{MHA}(\mathbf{F}_V^{(i)}, \mathbf{T}_w^{(i)}, \mathbf{T}_w^{(i)})}, \end{aligned} \quad (1)$$

$$\mathbf{F}_V^{(i+1)} = \text{DEL}^i(\mathbf{F}_V^{(i)}), \quad (2)$$

$$\underbrace{\text{softmax} \left( \frac{(Q\mathbf{W}^Q)(K\mathbf{W}^K)^T}{\sqrt{d/h}} \right) (\mathbf{V}\mathbf{W}^V)}_{\text{Concat}(\text{head}_1, \dots, \text{head}_h)W_V^O} \quad (3)$$

$$\hat{\mathbf{F}}_V = \mathbf{F}_{enc} + \overline{\text{MHA}}(\mathbf{F}_{enc}, \Psi(\mathbf{T}_s), \Psi(\mathbf{T}_s)),$$

where  $\theta_i$  is learnable parameters, MHA denotes the Multi-head Attention,  $W_V^O$  represents the linear projection matrix,  $\Psi(\cdot)$  is a MLP projection function,  $h$  is the number of head.

### Scale Adaptive Contextual Refinement

A severe challenge in UAV visual perception is the performance degradation in detecting small-scale objects. Although the Deformable DETR architecture bypasses the

traditional FPN, it has an inherent shortcoming. Specifically, the high-resolution feature maps, which are crucial for localizing small objects, possess a severely limited Effective Receptive Field. This results in a significant deficiency of contextual information, making it difficult for the model to distinguish small objects from complex background noise (Song et al. 2022; Wang et al. 2024). To address this, we insert a lightweight yet efficient module, named the Scale-Adaptive Contextual Refinement (SACR), between the encoder and decoder, as shown in Figure 4-B.

Specifically, we first employ parallel atrous convolutions with multiple distinct dilation rates on the highest-resolution feature map from the  $\hat{\mathbf{F}}_V = \{\hat{\mathbf{V}}^{(l)}\}_{l=1}^{L_s}$ , denoted as  $V_{ac}$ . It is capable of capturing rich, multi-scale contextual information without sacrificing spatial resolution. Formally:

$$V_{ac}^{(3)} = \text{Concat} \left( \text{Conv}_{1 \times 1}(\hat{\mathbf{V}}^{(3)}), \{\text{DConv}_{\{r_j\}}(\hat{\mathbf{V}}^{(3)})\}_{j=1}^M \right), \quad (4)$$

where  $\text{DConv}_{r_j}$  represents a  $3 \times 3$  atrous convolution,  $\{r_j\} = \{6, 12, 18\}$  denotes dilation rate.

After the contextual information is effectively aggregated, we perform an adaptive channel-wise feature recalibration to accentuate the feature channels crucial for small object recognition and suppress potential background noise. We capture local cross-channel interaction information via a one-dimensional convolution ( $\text{Conv}_k^{1D}$ ) with a kernel size of  $k$ , which is adaptively determined by a mapping function  $\psi$  based on the channel dimension  $C$ :

$$\begin{aligned} \mathbf{V}'^{(3)} &= \mathbf{w} \odot V_{ac}^{(3)}, \\ \mathbf{w} &= \sigma \left( \text{Conv}_k^{1D} (\text{GAP}(V_{ac})) \right), \\ k &= \left| \frac{\log_2(C) + b}{\gamma} \right|_{\text{odd}}, \end{aligned} \quad (5)$$

where  $\gamma$  and  $b$  are the hyperparameters and set to  $\gamma = 2$  and  $b = 1$ , respectively.  $|\cdot|_{\text{odd}}$  denotes the nearest odd integer. GAP is global average pooling, and  $\sigma$  is a Sigmoid function.

Method	In-domain Evaluation						Cross-domain Evaluation					
	HOTA	DetA	AssA	HOTA <sub>S</sub>	HOTA <sub>M</sub>	LocA	HOTA	DetA	AssA	HOTA <sub>S</sub>	HOTA <sub>M</sub>	LocA
MOTR-V2 <sub>CVPR 2023</sub>	19.51	11.57	33.13	21.67	19.11	<b>83.80</b>	21.70	13.85	34.13	23.85	24.85	83.43
TransRMOT <sub>CVPR 2023</sub>	23.54	13.18	42.24	27.21	24.05	83.47	26.86	15.21	47.66	24.47	25.43	83.65
TempRMOT <sub>arXiv 2024</sub>	26.24	13.06	<b>53.22</b>	28.14	23.77	80.41	27.58	13.46	<b>56.84</b>	23.74	27.67	83.06
CDRMT <sub>INFFUS 2025</sub>	25.81	14.66	45.69	27.49	25.80	83.13	26.68	16.21	44.11	26.98	25.20	83.08
MGLT <sub>TIM 2025</sub>	26.16	14.83	46.47	26.39	26.10	82.44	27.66	15.18	50.60	26.94	28.19	83.94
HETrack (Ours)	<b>31.46</b>	<b>21.57</b>	46.23	<b>34.37</b>	<b>31.12</b>	82.77	<b>31.60</b>	<b>21.35</b>	47.10	<b>27.53</b>	<b>31.93</b>	<b>83.98</b>

Table 2: Comparison with state-of-the-art methods on the in-domain and cross-domain test sets. The best results are in **bold**.

Finally, the refined multi-scale visual features  $\mathbf{F}'_v = \{\mathbf{V}'^{(l)}\}_{l=1}^{L_s}$ , refined by the encoder, are fed into the decoder with object queries to learn the target representation  $D_t$ . Furthermore, we employ a Semantic Guidance Module to perform semantically target-aware. Its process is as follows:

$$\begin{aligned} Q'_{\text{det}} &= Q_{\text{det}} + \text{CrossAttn}(Q_{\text{det}}, T_w, T_w), \\ D_t &= \text{Decoder}(\mathbf{F}'_v, \text{Concat}(Q_{\text{tra}}, Q'_{\text{det}})). \end{aligned} \quad (6)$$

## Loss Functions

To train the tracker, the loss is computed through a linear combination of four specialized loss terms:

$$\mathcal{L} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{L_1}\mathcal{L}_{L_1} + \lambda_{giou}\mathcal{L}_{giou} + \lambda_{ref}\mathcal{L}_{ref}. \quad (7)$$

where, the constituent losses  $\mathcal{L}_{cls}$ ,  $\mathcal{L}_{L_1}$ , and  $\mathcal{L}_{giou}$  correspond to the focal loss, L1 loss, and GIoU loss. Each term is scaled by a corresponding hyperparameter  $\lambda$ , which controls its relative importance during the training process.

## Experiments

### Implementation Details

The main architectural settings follow those in (Wu et al. 2023). The entire training is deployed on 8 NVIDIA A100 GPUs with a batch size of 1 for 100 epochs. We filtered the target bounding boxes by applying a score threshold of 0.5 and a referring matching score threshold  $\beta_{ref} = 0.4$ . AerialMind utilizes 63 sequences from the VisDrone for the training set and the remaining 17 sequences for in-domain testing. Additionally, we select 13 representative sequences from the UAVDT to serve as the cross-domain test set.

### Evaluation Metrics

To evaluate the overall tracking performance on AerialMind, we adopt the standard Higher Order Tracking Accuracy  $\text{HOTA} = \sqrt{\text{DetA} \cdot \text{AssA}}$  metric (Luiten et al. 2021). To facilitate deeper and more fine-grained diagnostic analysis of model performance, we introduce two attribute-based composite metrics:  $\text{HOTA}_S$  (Scene-Robustness) and  $\text{HOTA}_M$  (Motion-Resilience). For  $\text{HOTA}_S$ , the set of attributes  $\{A_i\}$  comprises Night, Occlusion, and Low Resolution; and the attributes of  $\text{HOTA}_M$  comprise Viewpoint Change, Scale Variation, Fast Motion, and Rotation. The general formula of attribute-based metrics is:  $\text{HOTA}_A = \sqrt[N]{\prod_{i=1}^N \text{HOTA}_{A_i}}$ ,  $N$  denotes the number of attributes included.

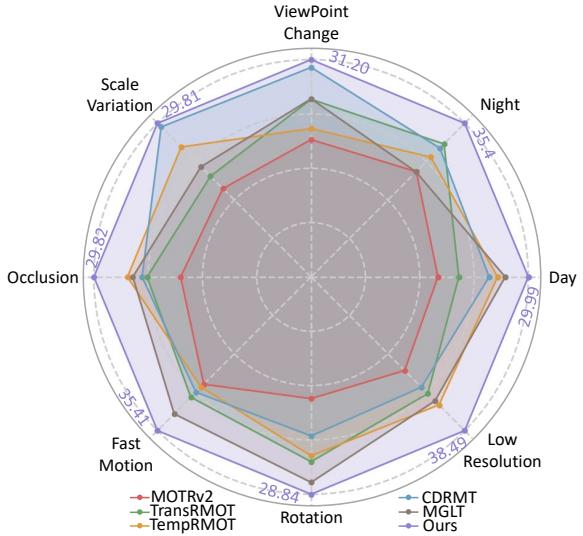


Figure 5: Comparison with state-of-the-art models in In-domain Evaluation with different attributes.

### Quantitative Results

We conduct extensive comparisons with state-of-the-art RMOT methods (MOTRV2 (Zhang, Wang, and Zhang 2023), TransRMOT (Wu et al. 2023), TempRMOT (Zhang et al. 2024), CDRMT (Liang et al. 2025), MGLT (Chen et al. 2025a)). The detailed results are presented in Table 2.

**In-domain Evaluation.** HETrack demonstrates state-of-the-art performance, achieving a HOTA score of 31.46%, significantly surpassing other leading end-to-end methods. Crucially, HETrack shows a pronounced advantage in our proposed attribute-based metric ( $\text{HOTA}_S$  34.37%,  $\text{HOTA}_M$  31.12%). This superiority is further substantiated by a detailed attribute-based analysis, as visualized in Fig 5. The results reveal that HETrack achieves the highest performance across all challenging attributes, and establishes a particularly significant lead in scenarios involving Low Resolution (38.49%), Fast Motion (35.41%), and Night (35.4%) conditions. While HETrack enhances the ability for localizing these small-scale objects to improve overall detection accuracy (DetA 21.57%), it leads to a marginal decrease in the average localization score (LocA 82.77%).

**Cross-domain Evaluation.** To rigorously assess model

TransRMOT	TempRMOT	CDRMT	SKTrack	HFF-Track	HETrack
31.00	35.04	31.99	35.29	<b>36.18</b>	35.40

Table 3: HOTA performance comparison of the Refer-KITTI-V2 dataset.

Components	HOTA	DetA	AssA
w/o CFE & SACR	26.41	16.43	42.80
w/o CFE	28.27	18.53	43.49
w/o SACR	29.89	19.86	45.34
HETrack (Ours)	<b>31.46</b>	<b>21.57</b>	<b>46.23</b>

Table 4: Ablation studies of different components in HETrack. “w/o” denotes components not used.

generalization, we evaluate models on the cross-domain test set. HETrack continues to outperform all other methods, not only achieving state-of-the-art results in core metrics like HOTA(31.60%), DetA(21.35%), and LocA(83.98%), but also attaining the highest scores in our proposed attribute-based metrics,  $HOTA_S$ (27.53%) and  $HOTA_M$ (31.93%).

An interesting phenomenon emerges from this evaluation: most methods, including HETrack, yield higher HOTA scores than their in-domain results. We posit that this counterintuitive result stems from the intrinsic disparity in scene complexity between the domains. Specifically, our training domain (VisDrone) features ten distinct object categories, fostering rich and complex semantic expressions. In contrast, the cross-domain test set (UAVDT) is predominantly limited to vehicle-only annotations, which significantly constrains the semantic space and simplifies the language grounding challenge. It also validates the distributional diversity and the value for pre-training of AerialMind.

**Ground-level Evaluation.** As shown in Table 3, we compare with state-of-the-art methods like SKTrack (Li et al. 2025b) and HFF-Track (Zhao et al. 2025) on the complex expressions ground-level referring dataset Refer-KITTI-V2. Our method also demonstrates competitive performance (35.40% HOTA). It validates that our method provides universal benefits for referring understanding.

## Qualitative Results

We visualize several representative examples in Figure 6. HETrack successfully achieves precise detection and tracking of referent objects according to the given expressions in various challenging UAV scenarios, including night illumination, complex spatial relationships, and small objects. Most notably, Figure 6-D fully demonstrates the model’s advanced reasoning capabilities for implicit descriptions.

## Ablation Studies

We systematically evaluate the contribution of our two main innovations (Table 4). Our HETrack model achieves a state-of-the-art HOTA score of 31.46%. When the SACR module is removed, the performance drops to 29.89%, underscoring its critical role in enhancing small object percep-

Fusion methods	HOTA	DetA	AssA
Concat	28.88	18.76	44.83
Add	30.39	19.95	46.65
Cross-Attn.( $T_s$ )	30.52	19.21	<b>48.82</b>
Ours	<b>31.46</b>	<b>21.57</b>	46.23

Table 5: Ablation studies of Semantic Guidance Module.



Figure 6: Qualitative examples on AerialMind. HETrack successfully tracks objects according to the expression.

tion. Removing the CFE module leads to a more significant performance degradation, with the HOTA score falling to 28.27%, which highlights the importance of our synergistic vision-language fusion strategy. In Table 5, we compare different fusion strategies like feature concatenation, addition, and cross-attention with sentence-level features ( $T_s$ ).

## Conclusion

In this work, we propose the first large-scale referring multi-object tracking dataset in UAV scenarios. It presents the unique challenges inherent in aerial viewpoints and introduces fine-grained attribute evaluation. Additionally, we develop a novel semi-automated collaborative agent-based framework that significantly enhances annotation efficiency and quality. Furthermore, we propose HETrack as a strong performance baseline. Extensive experiments validate the challenging nature of our dataset and the superior effectiveness of HETrack. We hope this work paves the way for future research in aerial language-guided perception.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62172246, in part by Shandong Taishan Scholar Young Expert Project and Excellent Young Scientists Fund of Shandong Provincial Natural Science Foundation under Grant ZR2024YQ071, and in part by the Fundamental Research Funds for the Central Universities under Grant 22CX06037A, and in part by the Youth Innovation and Technology Support Plan of Colleges and Universities in Shandong Province under Grant 2021K1062, in part by the Criminal Inspection Key Laboratory of Sichuan Province under Grant 2024YB01, and in part by the Fundamental Research Funds for the Central Universities through the Youth Program under Grant 22CX06037A.

## References

- Chen, J.; Lin, J.; Zhong, G.; Yao, Y.; and Li, Z. 2025a. Multi-granularity Localization Transformer with Collaborative Understanding for Referring Multi-Object Tracking. *IEEE Transactions on Instrumentation and Measurement*.
- Chen, S.; Yu, E.; Li, J.; and Tao, W. 2024a. Delving into the trajectory long-tail distribution for multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19341–19351.
- Chen, S.; Yu, E.; and Tao, W. 2025. Cross-view referring multi-object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Chen, S.; Yu, Y.; Yu, E.; and Tao, W. 2025b. ReaMOT: A Benchmark and Framework for Reasoning-based Multi-Object Tracking. *arXiv preprint arXiv:2505.20381*.
- Chen, W.; Jia, H.; Lai, S.; Wu, K.; Xiao, H.; Hu, L.; and Yue, Y. 2025c. Free-T2M: Frequency Enhanced Text-to-Motion Diffusion Model With Consistency Loss. *arXiv preprint arXiv:2501.18232*.
- Chen, W.; Xiao, H.; Zhang, E.; Hu, L.; Wang, L.; Liu, M.; and Chen, C. 2024b. Sato: Stable text-to-motion framework. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 6989–6997.
- Chen, W.; Yu, K.; Haozhe, J.; Yuan, K.; Huang, Z.; Tian, B.; Lai, S.; Xiao, H.; Zhang, E.; Wang, L.; et al. 2025d. ANT: Adaptive Neural Temporal-Aware Text-to-Motion Model. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 9852–9861.
- Ding, H.; Liu, C.; He, S.; Jiang, X.; Torr, P. H.; and Bai, S. 2023. MOSE: A new dataset for video object segmentation in complex scenes. In *Proceedings of the IEEE/CVF international conference on computer vision*, 20224–20234.
- Ding, H.; Liu, C.; He, S.; Ying, K.; Jiang, X.; Loy, C. C.; and Jiang, Y.-G. 2025a. MeViS: A multi-modal dataset for referring motion expression video segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Ding, H.; Liu, C.; Wang, S.; and Jiang, X. 2022a. VLT: Vision-language transformer and query generation for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6): 7900–7916.
- Ding, H.; Ying, K.; Liu, C.; He, S.; Jiang, X.; Jiang, Y.-G.; Torr, P. H.; and Bai, S. 2025b. MOSEv2: A More Challenging Dataset for Video Object Segmentation in Complex Scenes. *arXiv preprint arXiv:2508.05630*.
- Ding, Z.; Hui, T.; Huang, J.; Wei, X.; Han, J.; and Liu, S. 2022b. Language-bridged spatial-temporal interaction for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Du, D.; Qi, Y.; Yu, H.; Yang, Y.; Duan, K.; Li, G.; Zhang, W.; Huang, Q.; and Tian, Q. 2018. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European conference on computer vision (ECCV)*.
- Du, D.; Zhu, P.; Wen, L.; Bian, X.; Lin, H.; Hu, Q.; Peng, T.; Zheng, J.; Wang, X.; Zhang, Y.; et al. 2019. VisDrone-DET2019: The vision meets drone object detection in image challenge results. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*.
- Du, Y.; Lei, C.; Zhao, Z.; and Su, F. 2024. ikun: Speak to trackers without retraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Guan, R.; Jia, L.; Yao, S.; Yang, F.; Xu, S.; Purwanto, E.; Zhu, X.; Man, K. L.; Lim, E. G.; Smith, J.; et al. 2025a. Watervg: Waterway visual grounding based on text-guided vision and mmwave radar. *IEEE Transactions on Intelligent Transportation Systems*.
- Guan, R.; Liu, J.; Jia, L.; Zhao, H.; Yao, S.; Zhu, X.; Man, K. L.; Lim, E. G.; Smith, J.; and Yue, Y. 2024. NanoMVG: USV-centric low-power multi-task visual grounding based on prompt-guided camera and 4D mmWave radar. *arXiv preprint arXiv:2408.17207*.
- Guan, R.; Ouyang, N.; Xu, T.; Liang, S.; Dai, W.; Sun, Y.; Gao, S.; Lai, S.; Yao, S.; Hu, X.; et al. 2025b. Da Yu: Towards USV-Based Image Captioning for Waterway Surveillance and Scene Understanding. *arXiv preprint arXiv:2506.19288*.
- Guan, R.; Zhang, R.; Ouyang, N.; Liu, J.; Man, K. L.; Cai, X.; Xu, M.; Smith, J.; Lim, E. G.; Yue, Y.; et al. 2025c. Talk2radar: Bridging natural language with 4d mmwave radar for 3d referring expression comprehension. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 10884–10891. IEEE.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- Hui, T.; Huang, S.; Liu, S.; Ding, Z.; Li, G.; Wang, W.; Han, J.; and Wang, F. 2021. Collaborative spatial-temporal modeling for language-queried video actor segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Khoreva, A.; Rohrbach, A.; and Schiele, B. 2019. Video object segmentation with language referring expressions. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV 14*. Springer.
- Lai, S.; Hu, L.; Wang, J.; Berti-Equille, L.; and Wang, D. 2023. Faithful vision-language interpretation via concept

- bottleneck models. In *The Twelfth International Conference on Learning Representations*.
- Lai, S.; Liao, M.; Hu, Z.; Yang, J.; Chen, W.; Xiao, H.; Tang, J.; Liao, H.; and Yue, Y. 2025. Learning New Concepts, Remembering the Old: Continual Learning for Multimodal Concept Bottleneck Models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 12314–12322.
- Li, Y.; Liu, X.; Liu, L.; Fan, H.; and Zhang, L. 2025a. Lamot: Language-guided multi-object tracking. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 6816–6822. IEEE.
- Li, Y.; Zhou, S.; Qin, Z.; and Wang, L. 2025b. Visual-Linguistic Feature Alignment with Semantic and Kinematic Guidance for Referring Multi-Object Tracking. *IEEE Transactions on Multimedia*.
- Liang, C.; Wang, W.; Zhou, T.; Miao, J.; Luo, Y.; and Yang, Y. 2023. Local-global context aware transformer for language-guided video segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Liang, S.; Guan, R.; Lian, W.; Liu, D.; Sun, X.; Wu, D.; Yue, Y.; Ding, W.; and Xiong, H. 2025. Cognitive Disentanglement for Referring Multi-Object Tracking. *Information Fusion*.
- Liao, Y.; Liu, S.; Li, G.; Wang, F.; Chen, Y.; Qian, C.; and Li, B. 2020. A real-time cross-modality correlation filtering method for referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Lindeberg, T. 2013. *Scale-space theory in computer vision*, volume 256. Springer Science & Business Media.
- Liu, J.; Chen, Q.; Wang, Z.; Tang, Y.; Zhang, Y.; Yan, C.; Wang, D.; Li, X.; and Zhao, B. 2025. AerialVG: A Challenging Benchmark for Aerial Visual Grounding by Exploring Positional Relations. *arXiv preprint arXiv:2504.07836*.
- Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
- Luiten, J.; Osep, A.; Dendorfer, P.; Torr, P.; Geiger, A.; Leal-Taixé, L.; and Leibe, B. 2021. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*.
- Luo, G.; Zhou, Y.; Sun, X.; Cao, L.; Wu, C.; Deng, C.; and Ji, R. 2020. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*.
- Luo, R.; and Shakhnarovich, G. 2017. Comprehension-guided referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Ma, Z.; Yang, S.; Cui, Z.; Zhao, Z.; Su, F.; Liu, D.; and Wang, J. 2024. Mls-track: Multilevel semantic interaction in rmot. *arXiv preprint arXiv:2404.12031*.
- Seo, S.; Lee, J.-Y.; and Han, B. 2020. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*. Springer.
- Song, M.; Song, W.; Yang, G.; and Chen, C. 2022. Improving RGB-D salient object detection via modality-aware decoder. *IEEE Transactions on Image Processing*, 31: 6124–6138.
- Sun, P.; Cao, J.; Jiang, Y.; Zhang, R.; Xie, E.; Yuan, Z.; Wang, C.; and Luo, P. 2020. Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*.
- Sun, Z.; Liu, Y.; Zhu, H.; Gu, Y.; Zou, Y.; Liu, Z.; Xia, G.-S.; Du, B.; and Xu, Y. 2025. RefDrone: A Challenging Benchmark for Referring Expression Comprehension in Drone Scenes. *arXiv preprint arXiv:2502.00392*.
- Wang, G.; Chen, C.; Hao, A.; Qin, H.; and Fan, D.-P. 2024. Windb: hmd-free and distortion-free panoptic video fixation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Wang, X.; Yang, D.; Liao, Y.; Zheng, W.; Dai, B.; Li, H.; Liu, S.; et al. 2025. UAV-Flow Colosseo: A Real-World Benchmark for Flying-on-a-Word UAV Imitation Learning. *arXiv preprint arXiv:2505.15725*.
- Wu, D.; Dong, X.; Shao, L.; and Shen, J. 2022. Multi-level representation learning with semantic alignment for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Wu, D.; Han, W.; Wang, T.; Dong, X.; Zhang, X.; and Shen, J. 2023. Referring multi-object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.
- Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*. Springer.
- Zhang, Y.; Wang, T.; and Zhang, X. 2023. Motrv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.
- Zhang, Y.; Wu, D.; Han, W.; and Dong, X. 2024. Bootstrapping Referring Multi-Object Tracking. *arXiv preprint arXiv:2406.05039*.
- Zhao, Z.; Hao, Y.; Zhang, M.; Liu, Q.; Li, B.; Sui, D.; He, S.; and Chen, X. 2025. HFF-Tracker: A Hierarchical Fine-grained Fusion Tracker for Referring Multi-Object Tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Zhou, T.; Porikli, F.; Crandall, D. J.; Van Gool, L.; and Wang, W. 2022. A survey on deep learning technique for video segmentation. *IEEE transactions on pattern analysis and machine intelligence*.



Figure 7: Overview of the AerialMind, including complex aerial scenarios, large-scale annotations, and diverse evaluation.

## Appendix

The following sections present comprehensive examinations of our AerialMind dataset, detailed methodology analysis, and extensive experimental evaluations to provide deeper insights into our contributions. **Section A** offers additional statistical analysis and visualizations of the AerialMind dataset, demonstrating its richness and complexity across multiple dimensions. **Section B** presents comprehensive implementation details, including detailed training strategies and hyperparameter configurations. It provides extensive experimental results comparing our method with state-of-the-art approaches on both AerialMind and existing benchmarks, including ablation studies and additional qualitative results. Finally, **Section C** discusses the broader implications of our work, limitations, and potential directions for future research in aerial language-guided perception.

### A.AerialMind Benchmark

As shown in Figure 7, we demonstrate the diverse aerial scenarios that form the foundation of our benchmark. The collage of representative scenes, which range from complex urban intersections and highway interchanges to recreational facilities and commercial districts. It illustrates the unprecedented environmental diversity that distinguishes our dataset from existing ground-level benchmarks. As demonstrated in Table 6, our AerialMind dataset exhibits distinctive characteristics that establish it as a comprehensive benchmark for referring multi-object tracking in UAV Scenarios.

### Dataset Analysis and Statistics

AerialMind contains 48,485 frames, which has a moderate scale compared to some existing datasets (Wu et al. 2023; Chen et al. 2025a). It reflects our deliberate focus on scene diversity rather than temporal redundancy. Unlike Refer-Dance (Du et al. 2024) and CRMOT (Chen, Yu, and Tao

2025), which achieve high frame counts through repetitive indoor scenarios (dance studios/stages) or multi-viewpoint captures of limited environments, our dataset encompasses over 70 distinct aerial scenarios, providing richer environmental diversity and visual complexity.

Our dataset achieves 247.4 expressions per sequence, ranking second among all benchmarks, which reflects our emphasis on sophisticated semantic interactions rather than simplistic lexical permutations (Zhang et al. 2024). Our commitment is to capturing the nuanced linguistic complexity inherent in real-world aerial operations. Most significantly, AerialMind leads in both distinct expressions (7,601) and distinct instances (8,778), demonstrating unprecedented semantic richness and object diversity. This achievement validates our dataset’s capacity to encompass the full spectrum of aerial referring scenarios, from fine-grained object discrimination to complex multi-target reasoning tasks.

The temporal ratio per expression of 0.707 reveals another crucial design principle: our referring events are temporally dynamic rather than persistent throughout entire sequences. This characteristic reflects the realistic nature of aerial surveillance and monitoring tasks, where target objects frequently enter and exit the field of view due to UAV mobility and scene complexity. Unlike datasets with near-complete temporal coverage (e.g., Refer-KITTIV2 (Zhang et al. 2024) at 0.988), our design captures the intermittent and event-driven nature of real-world aerial operations.

As shown in Figure 8, these examples demonstrate several key innovations: **(1) Complex spatial understanding**, exemplified by expressions like “People who live next to entertainment facilities” and “Autos positioned on right of straight road”, which require understanding of complex spatial relationships from aerial perspectives; **(2) Fine-grained semantic discrimination**, such as distinguishing “Individuals traveling via electric bicycles” from regular bicycles, demanding precise object categorization capabilities; **(3) Dy-**

Dataset	Source	Frames	Expressions Per Sequence	Distinct Expressions	Distinct Instances	Temporal Ratio Per Expression
Refer-KITTI	CVPR <sub>2023</sub>	6074	49.7	215	637	0.550
Refer-Dance	CVPR <sub>2024</sub>	67304	30.5	48	650	0.939
Refer-BDD	IEEE TIM <sub>2025</sub>	4610	92.2	1212	6246	0.538
Refer-UE-City	arXiv <sub>2024</sub>	6207	51	—	—	0.780
Refer-KITTIV2	arXiv <sub>2024</sub>	7938	464.6	7193	740	0.988
CRTrack	AAAI <sub>2025</sub>	82338	43.5	132	—	—
LaMOT*	ICRA <sub>2025</sub>	27381	2.3	3	7889	—
AerialMind	Ours	48485	247.4	7601	8778	0.707

Table 6: Comparison of referring multi-object tracking datasets. LaMOT\* represents the UAV subset.

**namic motion understanding**, as shown in “Black automobiles traveling leftward to rightward”, which integrates color attributes with directional motion analysis; **(4) Logical reasoning expressions**, including “Cars invading other lanes”, which requires logical inference about traffic violations rather than simple visual matching; **(5) Comprehensive negative sampling**, the no-target scenarios like “The person walking on the crosswalk”.

This combination of environmental diversity, semantic sophistication, and temporal dynamics positions AerialMind as a uniquely challenging benchmark that mirrors the complexity demands of practical aerial intelligence systems.

## Evaluation Metrics

The design of our evaluation metrics framework is grounded in a fundamental understanding of the dual nature of challenges inherent in UAV-based referring multi-object tracking systems. Traditional RMOT (Wu et al. 2023; Ma et al. 2024) evaluation relies primarily on the standard HOTA metric (Luiten et al. 2021), which provides a balanced assessment of detection and association capabilities through the geometric mean  $\text{HOTA} = \sqrt{\text{DetA} \cdot \text{AssA}}$ . While HOTA offers valuable insights into overall system performance, it fails to capture the nuanced challenge taxonomy that distinguishes aerial platforms from ground-based systems. It limits our ability to conduct targeted diagnostic analysis and identify specific failure modes in complex UAV scenarios.

To address this limitation, we introduce a theoretically motivated decomposition of challenges into two fundamental categories that reflect the inherent characteristics of aerial operations. Our Scene-Robustness metric ( $\text{HOTA}_S$ ) encompasses attributes that are intrinsic to the environmental context itself. Specifically, Night illumination conditions, Occlusion scenarios, and Low Resolution imagery. These attributes represent static environmental factors that exist independently of UAV behavior and would persist regardless of platform mobility or operational dynamics.

Conversely, our Motion-Resilience metric ( $\text{HOTA}_M$ ) captures challenges that are directly correlated with UAV maneuverability and operational dynamics, including Viewpoint Change, Scale Variation, Fast Motion, and Camera Rotation. This categorization reflects the fundamen-

tal insight that UAV platforms introduce unique challenges through their mobility. Unlike static environmental challenges, these motion-induced factors are largely controllable through flight planning and operational procedures, making their separate evaluation crucial for understanding system limitations under different operational paradigms.

The mathematical formulation of our attributes evaluation metrics  $\text{HOTA}_A = \sqrt[N]{\prod_{i=1}^N \text{HOTA}_{A_i}}$  employs the geometric mean to ensure that performance degradation in any constituent attribute significantly impacts the composite metric, thereby preventing compensation effects where strong performance in one attribute masks critical weaknesses in another. This approach is rooted in the fundamental principle that a robust model must demonstrate consistent reliability across all relevant challenge dimensions. Because operational failure in any single aspect can compromise mission effectiveness in real-world deployments.

## B. Additional Experimental Results

### Implementation Details

Our model employs ResNet50 (He et al. 2016) as the visual backbone and RoBERTa (Liu et al. 2019) as the language encoder. Following established practices in referring multi-object tracking, the multi-scale feature maps extracted by the visual backbone undergo encoder-decoder processing for comprehensive cross-modal fusion and spatial relationship modeling. The remaining architectural configurations adhere to the framework established in (Wu et al. 2023).

The training process spans 100 epochs using the AdamW (Loshchilov and Hutter 2019) optimizer with an initial learning rate of  $1 \times 10^{-4}$ . The learning rate undergoes decay by a factor of 10 at the 40<sup>th</sup> epoch to ensure convergence stability. The loss function incorporates multiple weighted components:  $\lambda_{cls}$ ,  $\lambda_{L_1}$ ,  $\lambda_{giou}$ , and  $\lambda_{align}$  are configured to 2, 5, 2, and 2, respectively, to balance classification accuracy, localization precision, and alignment quality. To accommodate the diverse linguistic expressions characteristic of aerial scenarios, we initialize the framework with 300 object queries. To ensure reproducibility and fair comparison across different runs, we maintain a fixed random

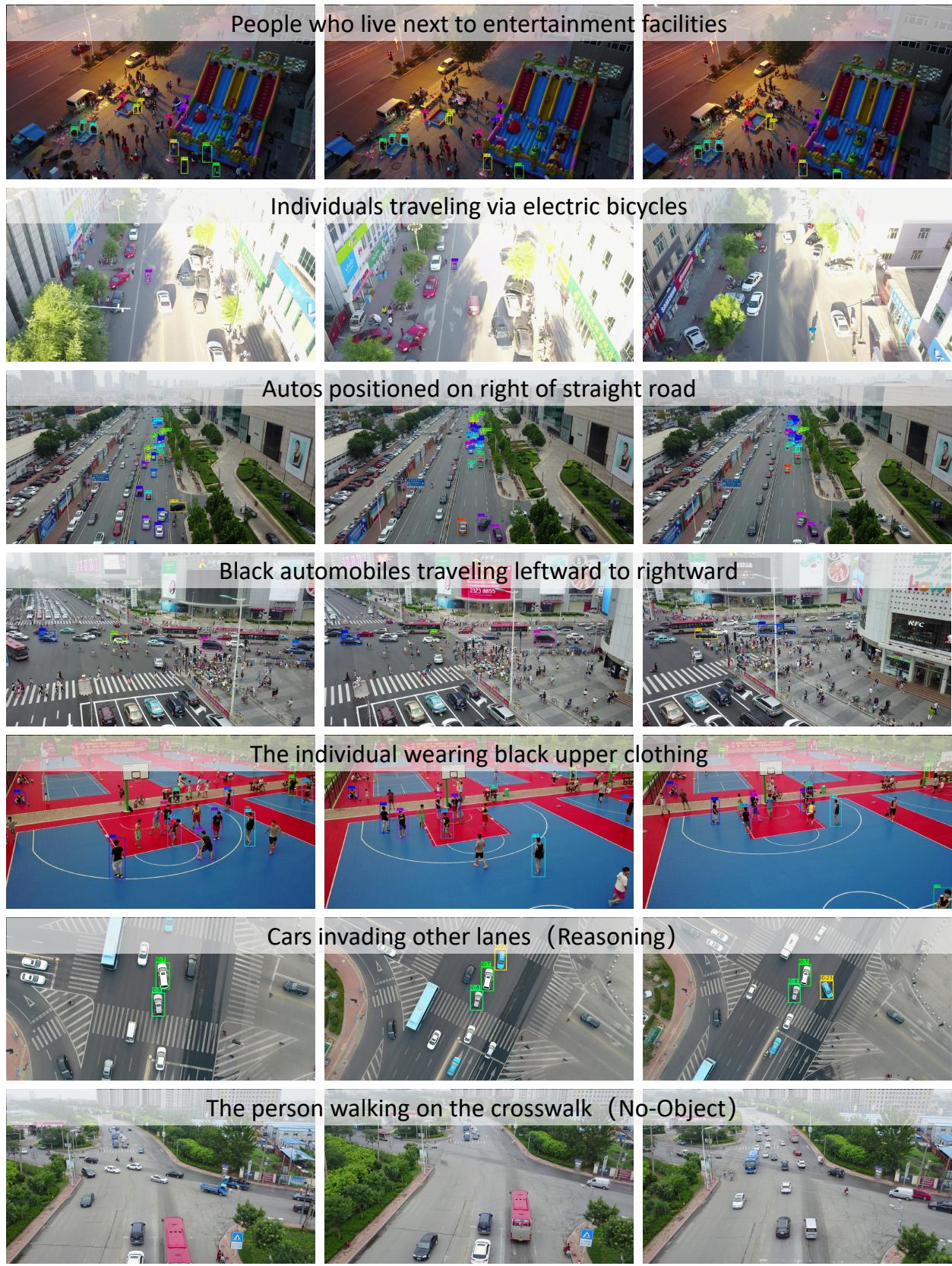


Figure 8: Representative examples from the AerialMind demonstrating diverse referring expressions and challenging scenarios.

Evaluation on Refer-KITTI-V2 Dataset									
Method	Source	HOTA	DetA	AssA	DetRe	DetPr	AssRe	AssPr	LocA
FairMOT	IJCV <sub>2021</sub>	22.53	15.80	32.82	20.60	37.03	36.21	71.94	78.25
ByteTrack	ECCV <sub>2022</sub>	24.59	16.78	36.63	22.60	36.18	41.00	69.63	78.00
TransRMOT	CVPR <sub>2023</sub>	31.00	19.40	49.68	36.41	28.97	54.59	82.29	89.82
iKUN	CVPR <sub>2024</sub>	10.32	2.17	49.77	2.36	19.75	58.48	68.64	74.56
TempRMOT*	arXiv <sub>2024</sub>	35.04	22.97	<b>53.58</b>	34.23	40.41	<b>59.50</b>	81.29	<u>90.07</u>
CDRMT	INFFUS <sub>2025</sub>	31.99	20.37	50.35	26.40	<b>46.26</b>	53.40	<b>85.90</b>	<b>90.36</b>
SKTrack	TMM <sub>2025</sub>	35.29	23.87	52.35	<u>39.97</u>	36.48	57.45	84.23	88.89
HFF-Track	AAAI <sub>2025</sub>	<b>36.18</b>	<u>24.64</u>	<u>53.27</u>	36.86	<u>41.83</u>	<u>59.42</u>	81.40	89.77
HETrack	Ours	35.40	<b>25.56</b>	49.18	<b>41.16</b>	39.53	53.50	<u>85.51</u>	89.80

Table 7: Comparison with state-of-the-art methods on the Refer-KITTI-V2 dataset. The best and second results are in **bold** and underline, respectively. \* indicates that the result was obtained by performing inference using the official open source code.

seed throughout both training and testing phases. Our HETrack model with 51.4M trainable parameters achieves 15.6 FPS on a single RTX 4080 during inference.

## Theoretical Foundations and Design Motivations

The architectural innovations in HETrack are grounded in rigorous theoretical foundations that address fundamental limitations in cross-modal representation learning and multi-scale object detection. This section elucidates the mathematical principles and theoretical motivations underlying our Co-evolutionary Fusion Encoder (CFE) and Scale Adaptive Contextual Refinement (SACR) modules.

**Co-evolutionary Fusion Encoder: Information-Theoretic Motivation.** Traditional fusion paradigms suffer from the information bottleneck problem (Lai et al. 2023, 2025), where cross-modal alignment is constrained by unidirectional information flow. Let  $V$  and  $L$  represent visual and linguistic feature distributions, respectively. Conventional approaches are limited by  $I(V; L) \leq \min(H(V), H(L))$ , where mutual information is bounded by the entropy of individual modalities. Our CFE addresses this fundamental limitation through bidirectional iterative. It enables visual and linguistic representations to evolve synergistically rather than independently.

**Scale Adaptive Contextual Refinement: Multi-scale Information Theory.** The SACR module is motivated by multi-scale information processing theory (Lindeberg 2013), which posits that optimal feature representation for small object detection requires principled information integration across multiple spatial scales. For UAV scenarios with extreme scale variations, the effective receptive ( $ERF_{optimal}$ ) field must satisfy:

$$ERF_{optimal} = \arg \max_r \sum_{s \in S} w_s \cdot MI(F_r^{(s)}, Y), \quad (8)$$

where  $F_r^{(s)}$  represents features at scale  $s$  with receptive field  $r$ , and  $MI$  denotes mutual information with target labels  $Y$ . Our atrous convolution configuration with dilation

rates  $\{6, 12, 18\}$  approximates this optimal multi-scale integration while preserving spatial resolution. The adaptive channel recalibration mechanism seeks to emphasize task-relevant features while reducing noise interference (Guan et al. 2025a) for robust representation.

## Quantitative Results

**Evaluation on Refer-KITTI-V2 Dataset** To further validate the generalizability and robustness of our proposed HETrack method, we conduct comprehensive evaluations on the Refer-KITTI-V2 dataset (Zhang et al. 2024). It represents the most extensive ground-perspective RMOT benchmark featuring complex linguistic descriptions. It's a crucial evaluation for our method's adaptability across diverse scenarios and semantic complexities.

As demonstrated in Table 7, HETrack achieves competitive performance with a HOTA score of 35.40%, ranking among the top-tier methods alongside SKTrack (35.29%) and HFF-Track (36.18%). Notably, our method demonstrates particular strengths in detection accuracy (DetA: 25.56%) and detection recall (DetRe: 41.16%), significantly outperforming recent methods such as CDRMT (DetA: 20.37%) and SKTrack (DetA: 23.87%). This superior detection capability directly validates the effectiveness of our Scale Adaptive Contextual Refinement (SACR) module, which, despite being originally designed for small object perception in aerial scenarios. It proves equally beneficial for ground-level object detection by enhancing contextual feature representation and reducing background noise. More significantly, our method achieves remarkable improvements in critical detection metrics, with DetRe reaching 41.16%—substantially higher than competing methods such as SKTrack (39.97%) and TempRMOT (34.23%). This enhancement in detection recall demonstrates the robust cross-modal alignment capabilities of our Co-evolutionary Fusion Encoder (CFE), which enables synergistic refinement of vision-language representations. The CFE's bidirectional fusion mechanism proves particularly effective in handling the complex linguistic expressions characteristic.

Attributes Evaluation on UAVDT test set									
Method	Source	Day	Night	ViewPoint Change	Scale Variation	Occlusion	Fast Motion	Rotation	Low Resolution
MOTRv2	CVPR <sub>2023</sub>	22.07	20.42	26.25	19.86	28.53	28.10	26.01	23.28
TransRMOT	CVPR <sub>2023</sub>	<u>28.90</u>	23.00	<u>31.61</u>	<u>22.28</u>	<u>28.59</u>	24.05	24.69	22.28
TempRMOT	arXiv <sub>2024</sub>	28.27	25.48	29.33	20.65	26.15	24.40	<b>39.64</b>	20.08
CDRMT	INFFUS <sub>2025</sub>	26.64	<u>26.78</u>	30.10	20.72	<b>29.94</b>	27.09	23.88	24.49
MGLT	IEEE TIM <sub>2025</sub>	28.39	25.62	31.57	19.41	26.68	<b>33.67</b>	30.64	<b>28.60</b>
HETrack	Ours	<b>32.18</b>	<b>30.16</b>	<b>33.57</b>	<b>26.23</b>	25.80	<u>31.10</u>	<u>37.96</u>	<u>26.82</u>

Table 8: Comparison of various models in Cross-domain Evaluation with different attributes. The best results are in **bold** and the second best are underlined.

These results provide compelling evidence of our method’s core innovations. They capture fundamental principles of referring multi-object tracking that generalize effectively across diverse visual domains. It demonstrates that HETrack constitutes a generalizable framework capable of adapting to different scene characteristics while maintaining its core strengths in cross-modal representation learning and contextual feature enhancement.

**Attributes evaluation on UAVDT** To provide a comprehensive assessment of our method’s cross-domain generalization capabilities, we present detailed attribute-based performance analysis on the UAVDT test set in Table 8.

HETrack’s consistent top-tier performance across seven out of eight challenging attributes, achieving four best and three second-best results. It provides compelling evidence of its superior cross-domain generalization capabilities (Table 8). Specifically, HETrack achieves state-of-the-art performance in the crucial attributes of Day, Night, ViewPoint Change, and Scale Variation. Furthermore, it demonstrates highly competitive results in Fast Motion, Rotation, and Low Resolution. These results confirm that our innovations successfully capture domain-agnostic tracking principles and are not overfit to the training domain’s data.

## Qualitative Results

To provide comprehensive insights into the practical effectiveness of our proposed HETrack method, we present detailed qualitative comparisons with state-of-the-art approaches TempRMOT (Zhang et al. 2024) and MGLT (Chen et al. 2025a) across challenging scenarios in Figure 9-11.

**In-domain Evaluation** As shown in Figure 9, the basketball court scene presents a particularly demanding test case involving dense crowds, frequent occlusions, and rapid multi-directional movements. Both TempRMOT (Zhang et al. 2024) and MGLT (Chen et al. 2025a) exhibit failures in detecting “individuals wearing dark upper garments” in this complex scene. In stark contrast, HETrack robustly detects and tracks all relevant individuals.

The second row showcases HETrack’s exceptional capability in handling the dual challenges of small object detection and partial occlusion by environmental elements. While

TempRMOT and MGLT fail to comprehensively identify automobiles along the driving path, missing numerous roadside vehicles and distant objects, HETrack demonstrates remarkable precision in detecting both prominently visible and partially occluded vehicles.

The nighttime intersection scene presents a more sophisticated challenge, requiring accurate interpretation of motion states rather than visual appearance matching. The referring expression “white automobiles in stationary condition” demands precise discrimination between stationary and moving vehicles of the same color category. TempRMOT (Zhang et al. 2024) demonstrates semantic confusion by incorrectly identifying moving white vehicles as targets, while MGLT (Chen et al. 2025a) suffers from incomplete detection, missing several legitimate stationary targets. HETrack achieves superior semantic precision by correctly identifying only the stationary white automobiles.

**Cross-domain Evaluation** The cross-domain evaluation on UAVDT (Figure 10) demonstrates HETrack’s superior generalization capabilities across diverse scenarios and challenging conditions. In the complex intersection scenario with “vehicles placed on the road surface,” both TempRMOT and MGLT exhibit significant detection failures, particularly missing small-scale vehicles in distant regions. HETrack demonstrates remarkable robustness, successfully detecting vehicles across various scales and maintaining stable associations throughout the temporal sequence.

The night “moving cars” scenario reveals critical limitations in existing methods’ motion understanding and low-light object detection capabilities. TempRMOT shows poor comprehension of motion states, failing to accurately distinguish moving vehicles from stationary ones, while both TempRMOT and MGLT struggle with detecting small-scale targets under challenging illumination conditions. In contrast, HETrack maintains comprehensive detection of small nighttime objects and demonstrates superior motion state interpretation, accurately identifying only the moving vehicles as specified by the referring expression.

**Ground-level Evaluation** To further validate the generalizability of our approach across different viewing perspectives, we evaluate HETrack on ground-level scenarios from

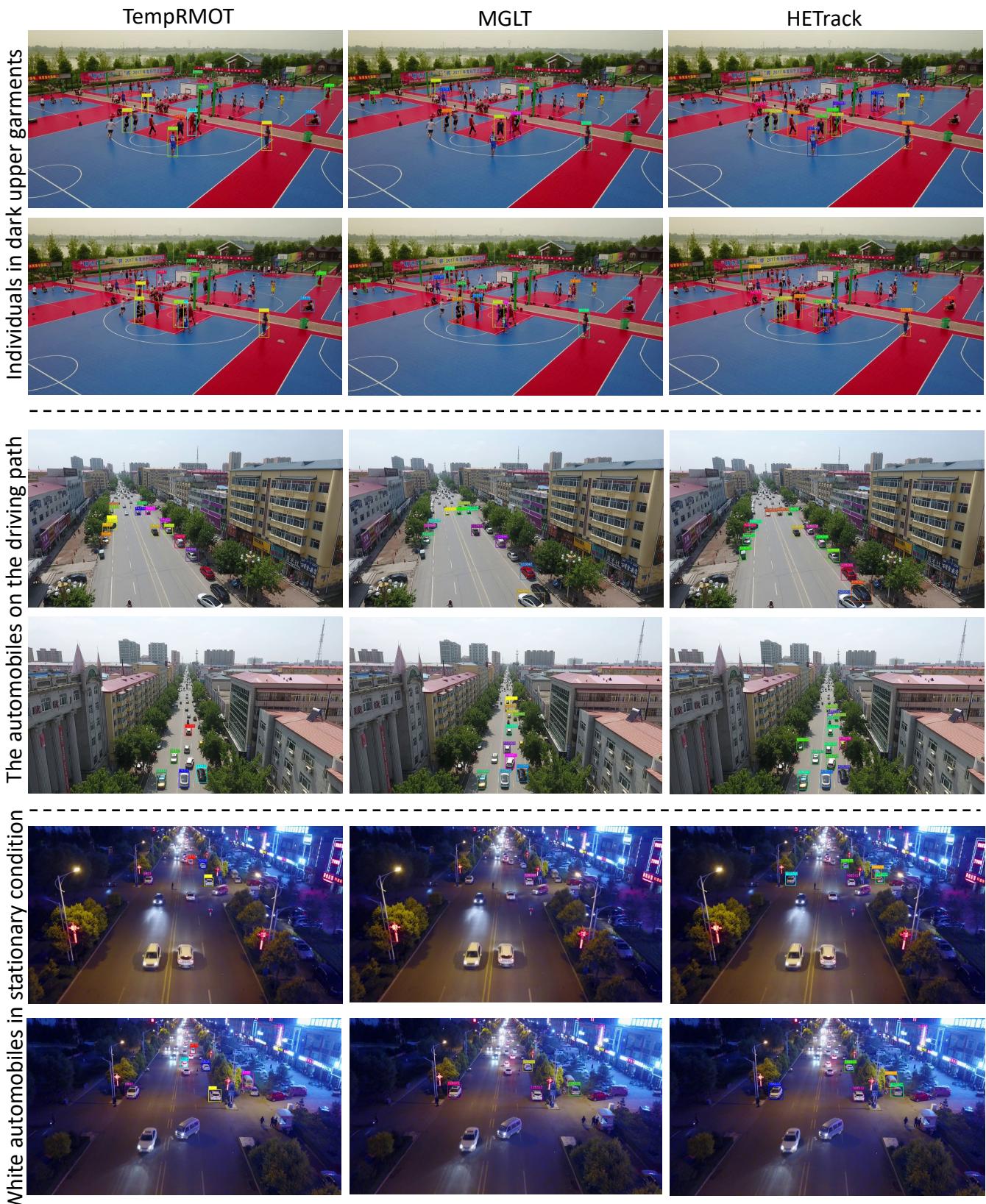


Figure 9: Qualitative comparison with the previous state-of-the-art methods on the In-domain evaluation of AerialMind.

The vehicles placed on the road surface



Moving cars



Figure 10: Qualitative comparison with the previous state-of-the-art methods on the Cross-domain evaluation of AerialMind.

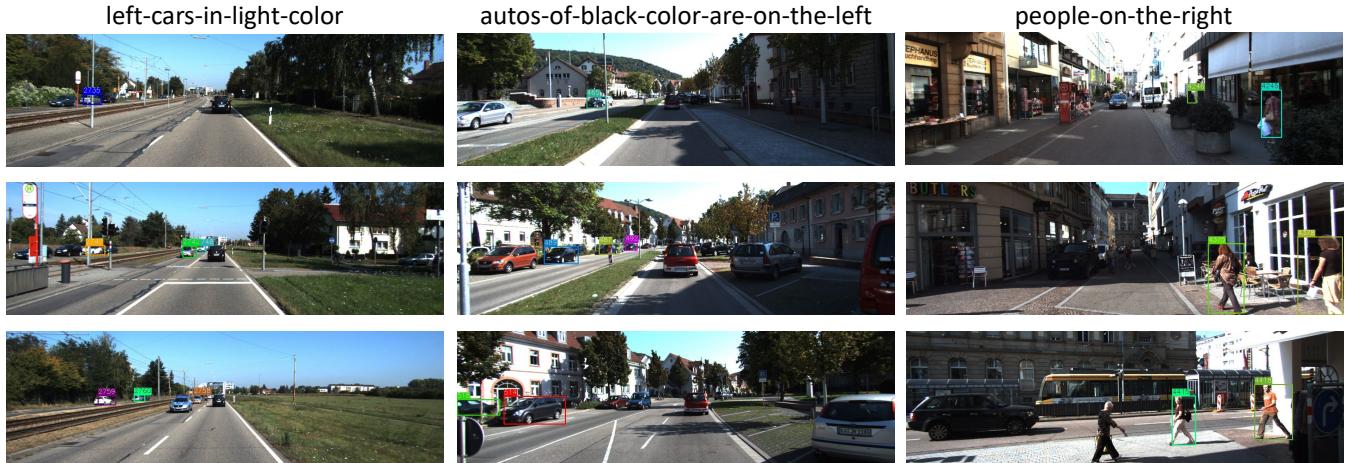


Figure 11: Qualitative results of our method on the Refer-KITTI-V2.

Method	HOTA	DetA	AssA
0.3	27.01	18.41	40.04
0.4	<b>31.46</b>	<b>21.57</b>	46.23
0.5	30.72	20.32	<b>46.83</b>
0.6	26.64	16.88	42.44

Table 9: Ablation studies on referring threshold  $\beta_{ref}$ .

the Refer-KITTI-V2 dataset (Figure 11). Despite being primarily designed for aerial scenarios, our method demonstrates remarkable adaptability to ground-perspective referring tasks. Across diverse urban environments and different viewpoint conditions, HETrack achieves precise object identification according to the given referring expressions.

### Ablation Study

**Referring Threshold** As demonstrated in Table 9, the model exhibits distinct performance patterns across the different thresholds. At the conservative threshold of  $\beta_{ref} = 0.3$ , it enables the model to capture more potential matches but introduces noise through false positive associations, as evidenced by the relatively lower DetA score of 18.41%. The optimal performance is achieved at  $\beta_{ref} = 0.4$ , where the model attains peak HOTA (31.46%) and DetA (21.57%) scores. This threshold represents an effective balance between detection sensitivity and association precision, enabling the model to maintain robust referring accuracy while maximizing overall tracking performance. Interestingly, increasing the threshold to  $\beta_{ref} = 0.5$  results in marginal degradation in overall performance (HOTA: 30.72%) despite achieving the highest association accuracy (AssA: 46.83%). This pattern indicates that while stricter thresholds improve the quality of established associations by filtering out ambiguous matches, they simultaneously reduce the system’s ability to detect valid referring instances, as reflected in the decreased DetA score (20.32%). The pronounced performance degradation at  $\beta_{ref} = 0.6$  (HOTA: 26.64%, DetA:

SACR Components	HOTA	DetA	AssA
Only Atrous Conv	29.70	19.43	45.81
Only Channel Recalibration	29.13	18.73	45.58
Full SACR (Ours)	<b>31.46</b>	<b>21.57</b>	<b>46.23</b>

Table 10: Ablation study on SACR module components.

16.88%, AssA: 42.44%) confirms that excessively strict thresholds severely limit the model’s detection capabilities. At this threshold, the system becomes overly selective, potentially missing legitimate referring instances due to the inherent uncertainty in cross-modal matching.

**Components of SACR module** Table 10 provides a systematic decomposition of the SACR module’s effectiveness, revealing the synergistic contributions of its constituent components for small object detection in aerial scenarios. When employing only the atrous convolution component, the model achieves moderate performance (HOTA: 29.70%, DetA: 19.43%), validating the importance of multi-scale contextual information capture for object localization. However, the limited performance improvement indicates that contextual information alone is insufficient for optimal object detection. The isolated channel recalibration mechanism yields relatively lower performance (HOTA: 29.13%, DetA: 18.73%), clearly suggesting that channel attention without adequate contextual support struggles to identify discriminative features for small-scale targets. The superior performance of our complete SACR module (HOTA: 31.46%, DetA: 21.57%) validates the synergistic integration of multi-scale contextual refinement and adaptive channel recalibration. It proves particularly valuable for effectively handling the diverse feature representations encountered across different scales in challenging UAV scenarios.

### C.Discussion

In this work, we introduced AerialMind, the first large-scale benchmark for Referring Multi-Object Tracking (RMOT) in UAV scenarios, alongside a strong baseline method, HETrack. Our contributions aim to bridge the significant gap between prevailing ground-level research and the unique, complex challenges posed by aerial platforms, thereby pushing the community towards developing more robust and versatile language-guided perception systems.

The primary significance of AerialMind lies in its establishment of a standardized evaluation platform for the aerial domain. By systematically incorporating challenges such as drastic scale variations, complex spatial relationships, and dynamic scenes, complemented by the first-of-its-kind attribute-level annotations in the RMOT field, our benchmark facilitates a more granular and insightful analysis of model capabilities. However, as a foundational step, we acknowledge its current limitations. The benchmark is constructed by extending existing public datasets, namely VisDrone and UAVDT. They inherited from these foundational datasets the presence of minor, pre-existing annotation errors. While we implemented a rigorous process to ensure the quality of our own annotations, it prevents the completely correct ground-truth labeling in a small subset of cases. Our future work will involve augmenting the benchmark with self-captured data, featuring a wider array of scenarios and more intricate object interactions.

While HETrack demonstrates competitive performance across various scenarios, several limitations warrant attention for future development. First, the current architecture relies on traditional vision-language fusion paradigms without leveraging the advanced reasoning capabilities of large language models (LLMs). Second, the computational overhead of our method presents challenges for real-time deployment on resource-constrained UAV platforms. The development of lightweight RMOT model variants represents a crucial direction for future research.