AerialMind: Towards Referring Multi-Object Tracking in UAV Scenarios

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

[1]Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China)
[2]Shandong Key Laboratory of Intelligent Oil & Gas Industrial Software
[3]Thrust of Artificial Intelligence, The Hong Kong University of Science and Technology (Guangzhou)
[4]Purple Mountain Laboratories
[5]School of Advanced Technology, Xi'an Jiaotong-Liverpool University
[6]School of Automation and Intelligent Sensing, Shanghai Jiao Tong University
[7]College of Science and Engineering, James Cook University
[8]Institute of Big Data, College of Computer Science and Artificial Intelligence, Fudan University
[9]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) 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.

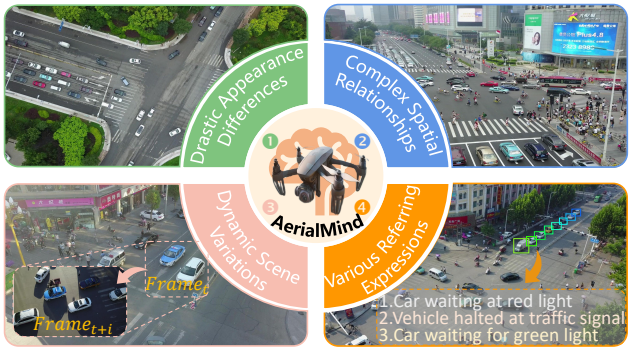

Figure 1: Overview of the challenges in AerialMind dataset.

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: ❶ Drastic Appearance Difference: Changes in flight altitude and viewpoints cause dramatic differences in object appearance; ❷ Complex Spatial Relationships: Object relationships under aerial view perspectives are more

*Corresponding author: runwayrwguan@hkust-gz.edu.cn

intricate; ❸ Dynamic Scene Variations: The high maneuverability of UAVs brings continuously changing scenes and illumination conditions; ❹ 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, 2025b,a), 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$_{17}$ (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	$CVPR_{2023}$	18	✗	✗	✗	818	49	10.7	8.8K	0.36M
Refer-Dance	$CVPR_{2024}$	65	✗	✗	✗	1.9K	25	0.34	650	0.55M
Refer-KITTI-V2	$arXiv_{2024}$	21	✗	✗	✗	9.8K	617	6.7	65.4K	3.06M
Refer-UE-City	$arXiv_{2024}$	12	✗	✗	✗	714	–	10.3	–	0.55M
Refer-BDD	IEEE TIM_{2025}	50	✗	✗	✗	4.6K	225	15.3	70.4K	1.50M
CRTrack	$AAAI_{2025}$	41	✓	✗	✗	344	43	–	–	0.79M
LaMOT*	IEEE $ICRA_{2025}$	62	✗	✗	✗	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.

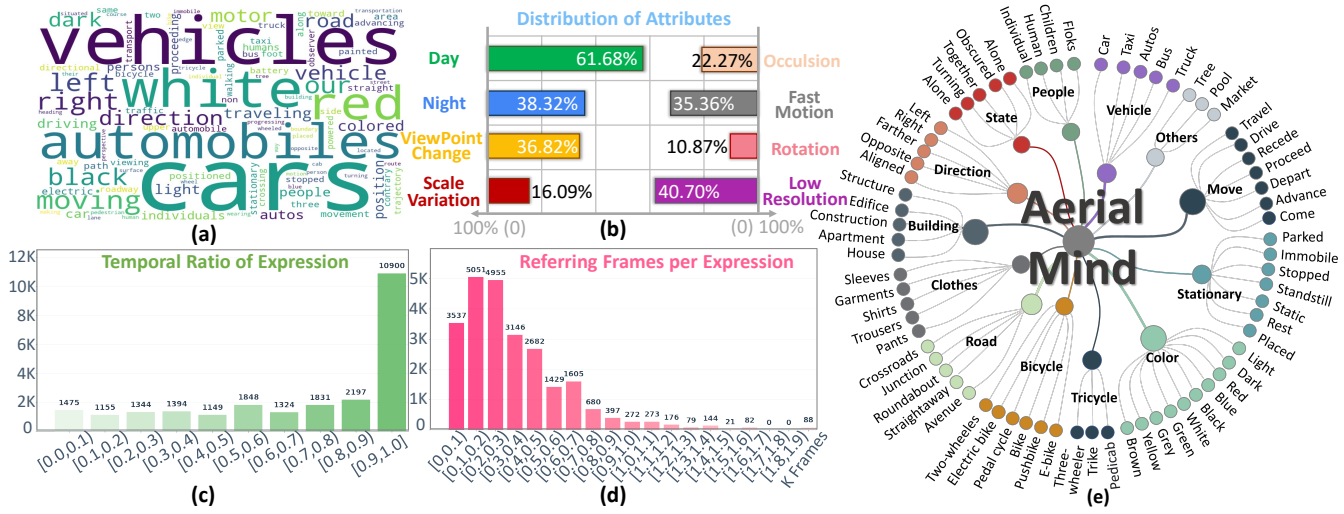

Figure 2: Overview of the AerialMind dataset statistics. It shows the distribution and diversity of (a) vocabulary, (b) challenging attributes, (c-d) temporal characteristics, and (e) semantic concepts.

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 $HOTA_S$ and $HOTA_M$, 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**ssistant (**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 balance 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

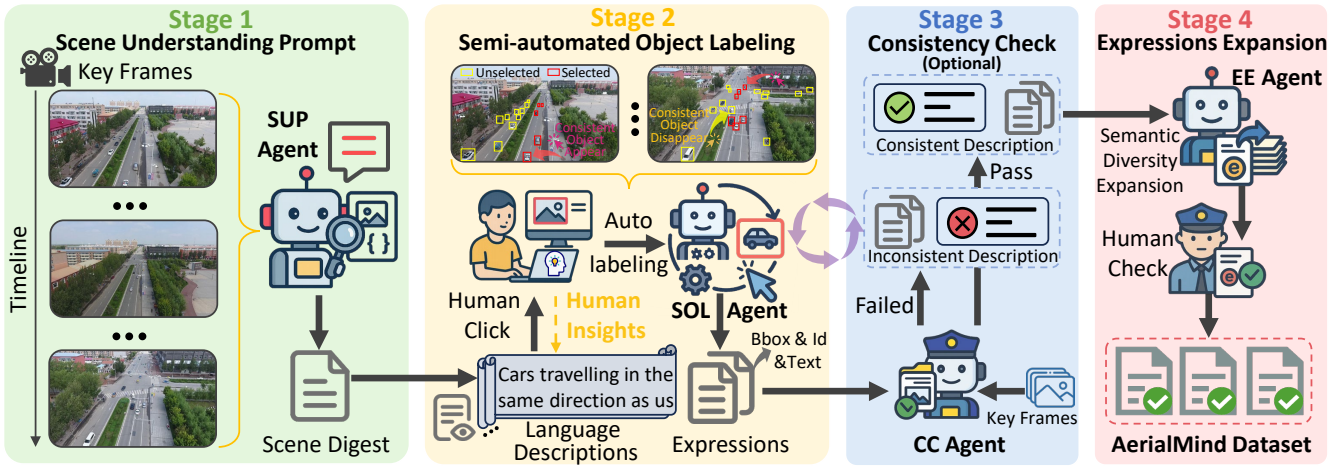

Figure 3: Overview of the four-stage annotation process in the COALA framework. This framework efficiently constructs the AerialMind dataset through multi-agent collaboration and human-computer interaction.

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

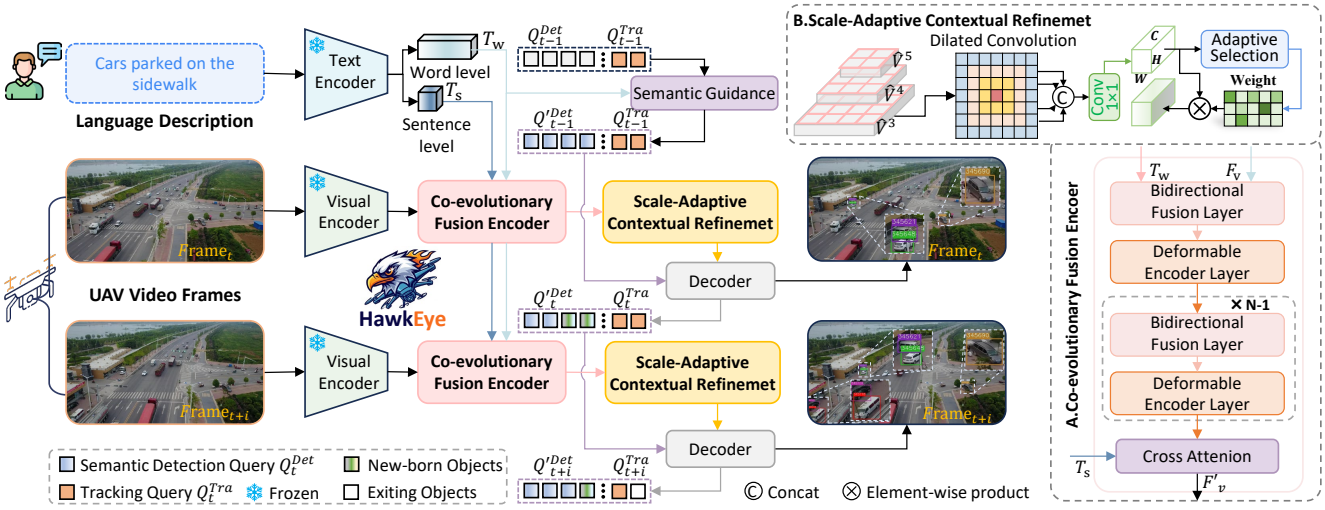

Figure 4: Overview of the HawkEyeTrack. Our key innovations include the Co-evolutionary Fusion Encoder for synergistic vision-language alignment and Scale-Adaptive Contextual Refinement for enhancing the perception of UAV scenarios.

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, $\mathbf{F}_{\text{enc}} = \mathbf{F'}_V^{(N_e)}$. 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:

$$\begin{aligned}\mathbf{F'}_V^{(i)}; \mathbf{T}_w^{(i)} &= \text{BFL}^i(\mathbf{F}_V^{(i)}, \mathbf{T}_w^{(i)}; \theta_i) \\ &= \underbrace{\mathbf{F}_V^{(i)} + \underbrace{\Delta\mathbf{F}_V^{(i)}}_{\Uparrow}; \mathbf{T}_w^{(i)} + \Delta\mathbf{T}_w^{(i)}}_{\text{MHA}(\mathbf{F}_V^{(i)}, \mathbf{T}_w^{(i)}, \mathbf{T}_w^{(i)})},\end{aligned} \tag{1}$$

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

$$\begin{aligned}&\underbrace{\text{softmax}\left(\frac{(Q\mathbf{W}^{\mathbf{Q}})(K\mathbf{W}^{\mathbf{K}})^T}{\sqrt{d/h}}\right)(V\mathbf{W}^{\mathbf{V}})}_{\Downarrow} \\ &\underbrace{\text{Concat}(\overline{\text{head}_1}, \ldots, \text{head}_h)W_V^O}_{\Downarrow} \\ &\hat{\mathbf{F}}_V = \mathbf{F}_{\text{enc}} + \overline{\text{MHA}}(\mathbf{F}_{\text{enc}}, \mathbf{\Psi}(\mathbf{T}_s), \mathbf{\Psi}(\mathbf{T}_s)),\end{aligned} \tag{3}$$

where θ_i is learnable parameters, MHA denotes the Multi-head Attention, W_V^O represents the linear projection matrix, $\mathbf{\Psi}(\cdot)$ 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:

$$V_{ac}^{(3)} = \text{Concat}\left(\text{Conv}_{1\times1}(\hat{\mathbf{V}}^{(3)}), \{\text{DConv}_{\{r_j\}}(\hat{\mathbf{V}}^{(3)})\}_{j=1}^{M}\right), \tag{4}$$

where DConv_{r_j} represents a 3×3 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 (Conv_k^{1D}) with a kernel size of k, which is adaptively determined by a mapping function ψ based on the channel dimension C:

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

where γ 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 σ is a Sigmoid function.

Method	In-domain Evaluation						Cross-domain Evaluation					
	HOTA	DetA	AssA	$HOTA_S$	$HOTA_M$	LocA	HOTA	DetA	AssA	$HOTA_S$	$HOTA_M$	LocA
$MOTR\text{-}V2_{CVPR\ 2023}$	19.51	11.57	33.13	21.67	19.11	**83.80**	21.70	13.85	34.13	23.85	24.85	83.43
$TransRMOT_{CVPR\ 2023}$	23.54	13.18	42.24	27.21	24.05	83.47	26.86	15.21	47.66	24.47	25.43	83.65
$TempRMOT_{arXiv\ 2024}$	26.24	13.06	**53.22**	28.14	23.77	80.41	27.58	13.46	**56.84**	23.74	27.67	83.06
$CDRMT_{INFFUS\ 2025}$	25.81	14.66	45.69	27.49	25.80	83.13	26.68	16.21	44.11	26.98	25.20	83.08
$MGLT_{TIM\ 2025}$	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)	**31.46**	**21.57**	46.23	**34.37**	**31.12**	82.77	**31.60**	**21.35**	47.10	**27.53**	**31.93**	**83.98**

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:

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

Loss Functions

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

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

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 λ, 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: $HOTA_S$ (Scene-Robustness) and $HOTA_M$ (Motion-Resilience). For $HOTA_S$, the set of attributes $\{A_i\}$ comprises Night, Occlusion, and Low Resolution; and the attributes of $HOTA_M$ 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.

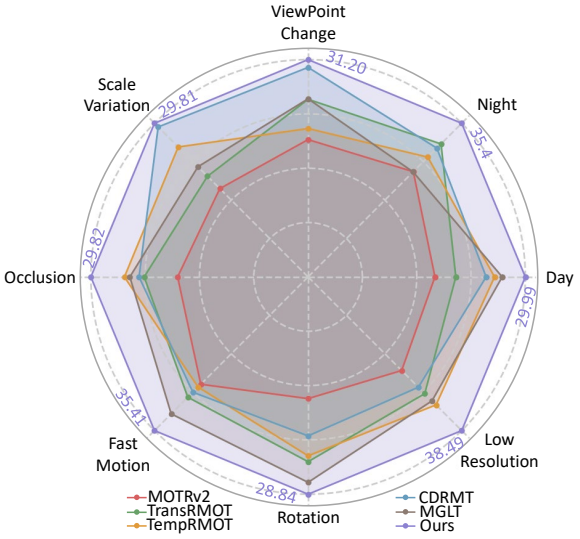

Figure 5: Comparison with state-of-the-art models in In-domain Evaluation with different attributes.

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 ($HOTA_S$ 34.37%, $HOTA_M$ 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	**36.18**	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)	**31.46**	**21.57**	**46.23**

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 perception. 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).

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	**48.82**
Ours	**31.46**	**21.57**	46.23

Table 5: Ablation studies of Semantic Guidance Module.

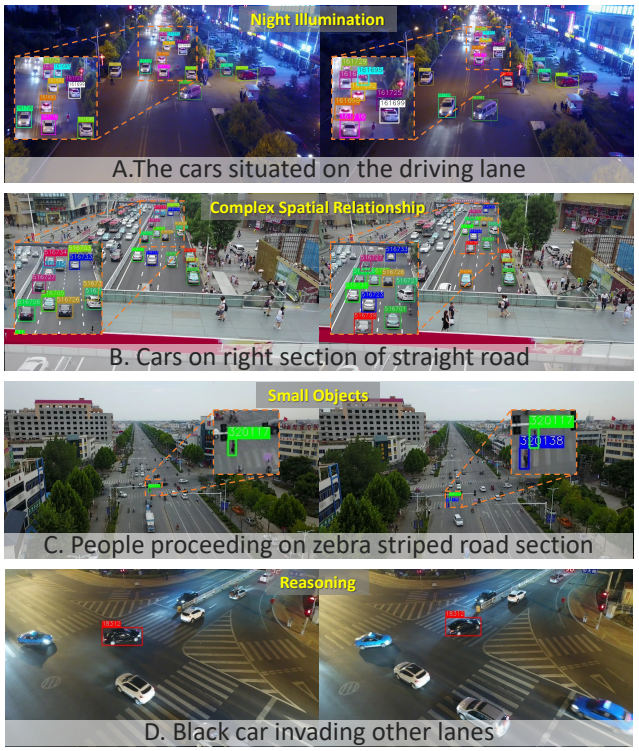

Figure 6: Qualitative examples on AerialMind. HETrack successfully tracks objects according to the expression.

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 muti-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.; 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. 2025a. 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. 2025b. 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.

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.

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*.

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*.

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*.