SAM2-OV: A Novel Detection-Only Tuning Paradigm for Open-Vocabulary Multi-Object Tracking

Yangkai Chen[1], Qiangqiang Wu[2], Guangyao Li[1], Junlong Gao[1], Guanglin Niu[3], Hanzi Wang[1]*

[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University, China
[2]City University of Hong Kong, Hong Kong
[3]School of Artificial Intelligence, Beihang University, China
yangkaichen@stu.xmu.edu.cn, qiangqwu2-c@my.cityu.edu.hk, liguangyao@stu.xmu.edu.cn,
jlgao@xmu.edu.cn, beihangngl@buaa.edu.cn, hanzi.wang@xmu.edu.cn

Abstract

Open-vocabulary multi-object tracking (OV-MOT) aims to track objects with unseen categories beyond the training set. While existing methods rely on pseudo video sequences synthesized from static images, they struggle to model realistic motion patterns, resulting in limited association performance in real-world scenarios. To alleviate these issues, we propose SAM2-OV, a novel association learning-free OV-MOT method that adopts a detection-only tuning paradigm, eliminating the need for synthetic sequences or spatiotemporal supervision and substantially reducing the overall learnable parameters. The core of our method is a Unified Detection Module (UDM), which effectively provides object-level prompts to enable SAM2 for OV-MOT. Enabled by UDM, SAM2-OV is the first to integrate SAM2 for OV-MOT, fully unleashing its zero-shot cross-frame association ability. To further enhance object association under occlusion and abrupt motion, we introduce a Motion Prior Assistance Module (MPAM) that incorporates motion cues into the mask selection process. In addition, a Semantic Enhancement Adapter (SEA) distilled from CLIP is used to improve classification generalization. A sparse prompting strategy is also adopted to reduce computational redundancy by triggering detection only on selected keyframes. As only the detection module is tuned on static images, the overall training process remains simple and efficient. Experiments on the TAO dataset demonstrate that SAM2-OV achieves state-of-the-art performance under the TETA metric, particularly on novel categories. Evaluations on the KITTI dataset show the strong zero-shot cross-domain transferability of our SAM2-OV.

Code — https://github.com/kaiycv/SAM2-OV.git.

Introduction

Multi-Object Tracking (MOT) is a fundamental task in computer vision, with broad applications in autonomous driving (Wu et al. 2021; Wu, Wan, and Chan 2021), video surveillance (Yi et al. 2024), and human–computer interaction (Zhuang et al. 2025; Li et al. 2025a). It involves continuously tracking multiple objects across video sequences, requiring both accurate recognition and consistent identity association. Most existing MOT methods (Tang et al. 2024; Luo et al. 2024) are trained on closed-set datasets with limited object categories, which severely hinders their ability to generalize to open-world scenarios with novel object categories during the testing stage.

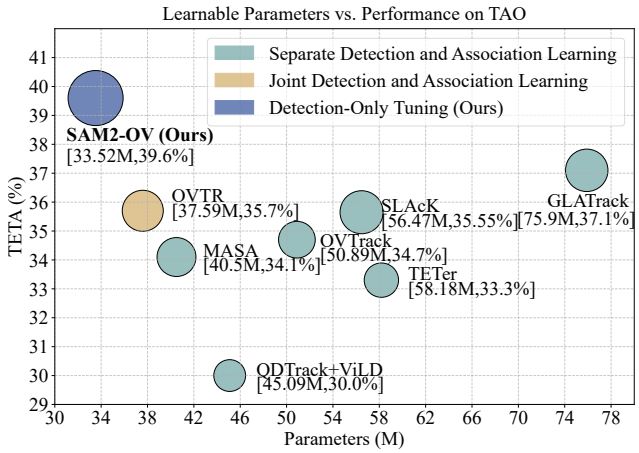

Figure 1: Comparison of learnable parameters and performance of OV-MOT methods on the TAO validation set. The horizontal axis indicates learnable parameters, the vertical axis shows TETA, and circle size reflects AssocA. SAM2-OV achieves superior performance with fewer parameters.

To bridge the gap between traditional MOT methods and real-world applications, Open-Vocabulary Multi-Object Tracking (OV-MOT) has been proposed, aiming to lift category constraints and enable continuous tracking of arbitrary objects. Most existing OV-MOT methods follow the separate detection and association learning paradigm (Li et al. 2023, 2024c), as illustrated in Figure 2a, and largely benefit from advances in open-vocabulary detection (OVD) (Du et al. 2022; Lin et al. 2022; Zhou et al. 2022; Fang, Pang, and Bai 2024; Ma et al. 2025).

For the joint detection and association learning paradigm, as illustrated in Figure 2b, the representative method is OVTR (Li et al. 2025b). Despite significant advances in

*Corresponding author.

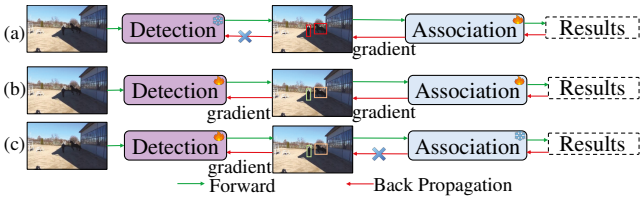

Figure 2: Comparison of three different OV-MOT paradigms: (a) illustrates the separate detection and association learning paradigm, matching detector-generated proposals using appearance and auxiliary cues. (b) shows the joint detection and association learning paradigm without explicitly modeling visual matching. (c) presents our detection-only tuning paradigm, which decouples association from supervision and presents a simplified learning paradigm.

these approaches, several key challenges remain unresolved: (1) Due to the tight coupling between location, classification, and association, most methods (Li et al. 2023, 2025b) adopt diffusion models to synthesize pseudo sequences from static images for joint training, but the generated frames fail to reflect real-world motion patterns, leading to suboptimal association performance. (2) Numerous proposals are required per frame to align with text, causing significant computational redundancy and higher costs in OV modeling.

To address the aforementioned challenges, we propose SAM2-OV, a novel open-vocabulary tracking method (Figure 2c) that adopts a detection-only tuning paradigm, eliminating the need for complex association learning. We leverage SAM2 (Ravi et al. 2024), a zero-shot video segmentation model with memory-based association, as the identity association module to decouple tracking from supervision. However, SAM2 is originally designed for general video object segmentation and directly applying SAM2 to OV-MOT is non-trivial, as it relies on human interaction for prompting and lacks the ability to autonomously localize and classify instance under open-vocabulary settings. To bridge this gap, we introduce a Unified Detection Module (UDM), which performs novel object detection and generates instance-level prompts to guide SAM2 for effective OV association. Enabled by UDM, SAM2-OV is the first to adapt SAM2 for OV-MOT to our knowledge, fully unleashing its zero-shot association potential in open-world scenarios.

Unlike previous methods (Gao and Wang 2023; Zhang et al. 2022; Li et al. 2023, 2025b) that rely on per-frame detection or dense proposal generation, SAM2-OV adopts a streamlined approach to unify detection and tracking. We introduce a lightweight Unified Detection Module (UDM) for generating bounding-box prompts on keyframes and assigning category and identity labels. To enhance classification generalization, a Semantic Enhancement Adapter (SEA) is integrated into UDM, aligning SAM2's multiscale features with general-purpose visual embeddings through distillation. To address SAM2's lack of motion modeling, the Motion Prior Assistance Module (MPAM) is introduced, incorporating motion priors to improve mask selection under occlusion and abrupt motion. As shown in Figure 1, our method achieves top open-vocabulary performance and stable tracking with minimal parameters for OV modeling. Additionally, a sparse detection strategy detects only on keyframes, propagating masks to intermediate frames via prompt-based association, reducing computational cost. Our main contributions are summarized as follows:

- We propose SAM2-OV, a detection-only tuning approach that eliminates complex association learning in OV-MOT, enabling stable long-term tracking across diverse categories with reduced reliance on per-frame detection.
- We propose a UDM for adaptive prompt generation to enable SAM2 for OV-MOT, and introduce SEA to enhance semantic generalization. Enabled by UDM, SAM2-OV is the first to explore SAM2 for OV-MOT, fully unleashing its zero-shot object association ability.
- We introduce MPAM that incorporates motion cues into the mask selection process of SAM2, enhancing temporal consistency and enabling more reliable identity association in complex dynamic scenarios.
- We evaluate SAM2-OV on the TAO and KITTI datasets, setting new state-of-the-art results on TAO. Specifically, our method outperforms SLAcK by 21.9% and 21.0% in TETA on the TAO validation and test sets for novel categories, respectively. It also surpasses OVTR by 9.3% in MOTA under zero-shot transfer on KITTI.

Related Work

Traditional MOT

Traditional multi-object tracking is mainly categorized into Tracking-by-Detection and joint detection–association frameworks. The former detects instances per frame and links them through matching strategies, as in SORT (Bewley et al. 2016), DeepSORT (Wojke et al. 2017), and OC-SORT (Cao et al. 2023). The latter jointly handles detection and identity association, represented by FairMOT (Zhang et al. 2021) and transformer-based MOTR (Carion et al. 2020). With the introduction of SAM2 (Ravi et al. 2024), recent works (Jiang et al. 2025; Vo et al. 2025; Videnovic, Lukezic, and Kristan 2025; Zeng, Huang, and Pei 2025; Ma et al. 2024; Song et al. 2025; Meinhardt et al. 2022) explore leveraging its cross-frame correspondence capabilities. However, these methods typically assume closed-set categories and struggle with unseen classes. To overcome this limitation, we introduce UDM, an open-vocabulary detection module that converts text descriptions into box prompts, enabling SAM2 to track instances beyond predefined categories.

Open-Vocabulary MOT

Open-vocabulary MOT seeks to track arbitrary categories through image-text alignment. OVTrack (Li et al. 2023) first establishes this task with CLIP-based localization and a track head. MASA (Li et al. 2024b) integrates SAM to provide unified segmentation-tracking, while SLAck (Li et al. 2024c) enhances association with semantic and spatial cues. OVTR (Li et al. 2025b) is the first end-to-end pipeline incorporating text features into transformer-based decoding.

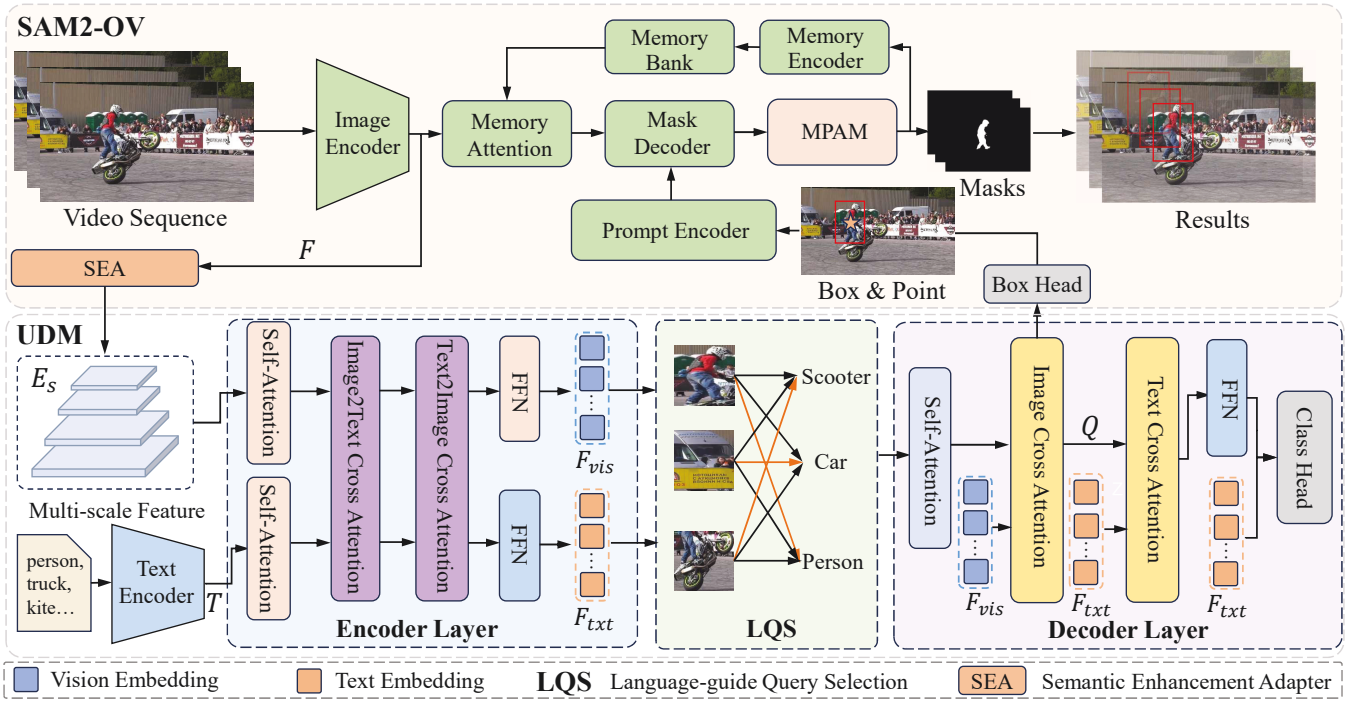

Figure 3: Overview of SAM2-OV. The top part shows the long-term tracking module, which combines SAM2's native components with MPAM, while the bottom part illustrates the detection module (UDM). UDM performs sparse detection on keyframes using SAM2 image features and CLIP text features to generate prompts with identity and category labels, which guide the long-term tracker for continued association. The entire framework follows a detection-only training paradigm.

Despite recent progress, many existing OV-MOT approaches (Li et al. 2024a; Liu et al. 2022) train association from pseudo-temporal sequences synthesized from static images, which limits their applicability in real-world dynamics. In contrast, our method employs a detection-only paradigm: it requires no explicit association training yet still maintains consistent identities across frames, offering improved robustness in complex scenarios.

Methodology

In this section, we first revisit existing OV-MOT learning paradigms and the paradigm our method builds upon, followed by an introduction to our overall framework, key components, and the training and inference strategies.

Preliminary

Revisit of previous OV-MOT paradigms. Most OV-MOT methods adopt a joint training paradigm that couples detection, classification, and association. Some leverage open-vocabulary detectors for per-frame recognition and match identities via appearance or semantics. Others directly predict detection and association without explicit matching. To simulate temporal context, they build pseudo video sequences S_{pseudo} via diffusion augmentation from static images, and jointly train detection and association modules D^*, A^* with the loss $\mathcal{L}_{overall}$ obtained by using the learning objective function $\mathcal{F}$:

$$\mathcal{L}_{overall} = \mathcal{F}(S_{pseudo}, D^*, A^*). \quad (1)$$

However, such paradigms incur high computational cost and suffer from domain gaps due to unrealistic motion, limiting real-world association performance.

Our detection-only tuning paradigm. To address the limitation of pseudo video sequences, we propose a detection-only tuning paradigm, eliminating the need for complex association learning with pseudo videos. Our paradigm allows the association module A to remain fixed and only trains the detection module D^* on static images I_{static}:

$$\mathcal{L}_{overall} = \mathcal{F}(I_{static}, D^*, A). \quad (2)$$

This decoupled training design avoids domain gaps and simplifies the pipeline while retaining robust identity tracking.

Overview

The overall pipeline of SAM2-OV is shown in Figure 3, consisting of two key modules: the SAM2-based association module and our Unified Detection Module (UDM). UDM predicts boxes, categories, and IDs to prompt SAM2's instance-aware processing for open-vocabulary tracking. SAM2-OV removes complex association learning by relying on SAM2's zero-shot cross-frame matching and training only the detection module. The following sections describe the proposed components and the training/inference setup.

Unified Detection Module

Pipeline. Given a key frame, UDM identifies novel-category objects by processing multi-scale features F extracted from

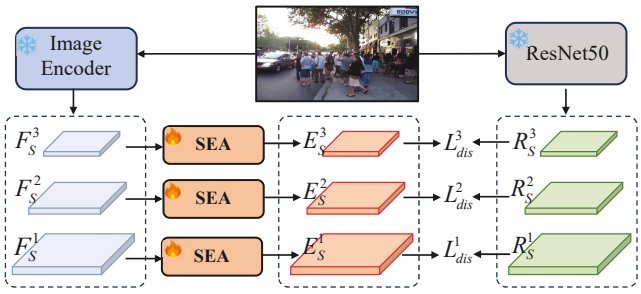

Figure 4: Overview of the proposed SEA distillation process designed to enhance open-vocabulary classification.

the SAM2 backbone. These features are first semantically enriched by SEA to enhance open-vocabulary recognition, and then refined through a cross-attention module that aligns visual and textual information, yielding F_{vis} and F_{txt}. To focus on the most semantically relevant content, the Language-Guided Query Selection module selects high-affinity tokens between F_{vis} and F_{txt} as input queries for the decoder. The decoder produces visual embeddings for bounding box regression and semantic embeddings for contrastive classification. The resulting boxes are used to prompt SAM2 for instance tracking, while the classification outputs assign category labels to each trajectory.

Architecture. To implement this process, UDM adopts a DETR-like (Zhu et al. 2020) encoder-decoder architecture that processes multi-scale features in a hierarchical manner. It jointly performs object localization and classification through query-based interaction guided by text-informed queries, enabling open-vocabulary generalization. Since the SAM2 backbone lacks strong semantic discriminability, we incorporate lightweight SEA modules at each scale to project features into a more classification-oriented space. Formally, let $F = \{F_s^1, F_s^2, \ldots, F_s^L\}$ be the extracted multi-scale features of the SAM2's backbone, where $F_s^l \in \mathbb{R}^{C_l \times H_l \times W_l}$ represents the feature map at the l-th scale. For each scale-specific feature map F_s^l, we introduce a corresponding adapter module $\text{SEA}(\cdot)$ that transforms it into a new feature space:

$$E_s^l = \text{SEA}_l(F_s^l) = \text{BN}(\text{Conv}_{1\times1}(F_s^l) + \phi(F_s^l)), \quad (3)$$

where $\text{Conv}_{1\times1}(\cdot)$ is a channel mapping, $\phi(\cdot)$ is a residual projection, and $\text{BN}(\cdot)$ denotes Batch Normalization. The resulting E_s^l provides adapted features for category-aware classification and is further refined through the distillation loss defined in Eq. 7.

For the detection module, we adopt an encoder-decoder architecture inspired by Grounding DINO (Liu et al. 2024) to strengthen semantic interaction between visual and textual modalities. The encoder leverages a bidirectional enhancement mechanism between image and text features, which improves the semantic alignment of visual representations and enhances the model's ability to perceive open-vocabulary concepts. Formally, let the text input be a sequence of features $T \in \mathbb{R}^{N\times d}$ extracted by CLIP's text encoder, where N is the number of tokens and d is the feature dimension. The visual input consists of the SEA-processed features $E = \{E_s^1, E_s^2, \ldots, E_s^L\}$. Within the encoder, cross-attention is applied to mutually enrich E and T in context, yielding joint semantically coherent representations.

The decoder adopts a query-based interaction mechanism to extract region-specific representations from text-guided queries. To further improve alignment between the detection outputs and the CLIP embedding space, we introduce an additional processing branch on the decoder's visual outputs. Let the decoder generate a set of object query features $Q = \{q_1, q_2, \ldots, q_K | q_i \in R^d\}$. Each query feature q_i is passed through two parallel branches. The first is a regression head that predicts bounding boxes via a simple MLP:

$$\widehat{b}_i = \text{Sigmoid}(\text{MLP}_{\text{box}}(q_i)), \widehat{b}_i \in \mathbb{R}^4, \quad (4)$$

and the second is an alignment head that projects q_i into the CLIP visual embedding space:

$$\widetilde{v}_i = \text{Normalize}(\text{MLP}_{\text{align}}(q_i)), \widetilde{v}_i \in \mathbb{R}^{d_{clip}}. \quad (5)$$

To perform category recognition in open-vocabulary scenarios, we introduce a non-parametric contrastive classification module that leverages decoder queries and encoder text features. The visual queries Q are refined via a text-guided cross-attention module and a feedforward network to produce semantically aligned object representations. These are then multiplied with category-level text features from the encoder to obtain classification logits $\mathcal{Z}$.

Motion Prior Assistance Module

Although SAM2 provides strong cross-frame matching, its mask selection lacks motion awareness, leading to ID switches under occlusion or abrupt viewpoint changes. To mitigate this, we introduce MPAM, which incorporates a lightweight Kalman filter, inspired by previous works (Bewley et al. 2016; Vo et al. 2025), to estimate object motion and guide mask selection. For each tracked object, the filter predicts the current position $\hat{b}_{\text{kf}}$, and computes IoUs with three candidate boxes $\{b_1, b_2, b_3\}$, producing motion-based scores $\{\text{IoU}_1^{\text{kf}}, \text{IoU}_2^{\text{kf}}, \text{IoU}_3^{\text{kf}}\}$. Each candidate's final score is then obtained by fusing motion and appearance cues via a weighted combination:

$$\text{Score}_i = \alpha \cdot \text{IoU}_i^{\text{kf}} + (1-\alpha) \cdot \text{IoU}_i^{\text{mask}}, \quad i \in \{1,2,3\}, \quad (6)$$

where α denotes the weight of the Kalman filter prediction in the mask selection process. For each target, the candidate mask with the highest fused score is selected as the final output for the current frame.

To avoid unstable motion prediction in early tracking, we use a frame threshold T_{w}, activating motion guidance only after a target is tracked for more than T_{w} frames. Before that, mask selection relies solely on SAM2's IoU predictions.

Training and Inference

Training Loss. Since our method adopts a detection-only training paradigm, the optimization involves only localization and classification in static images. Beyond direct supervision on classification outputs, we further enhance the semantic features of key modules to improve generalization.

To further enhance the representational capacity of SEA, we introduce a feature distillation mechanism that guides the SEA outputs to align with features extracted from a pre-trained ResNet-50, as illustrated in Figure 4. Let R_s^l denote the ResNet-50 feature at the corresponding scale. We employ the mean squared error loss for distillation, defined as:

$$\mathcal{L}_{\text{SEA}} = \frac{1}{L}\sum_{l=1}^{L} L_{dis}^{l} = \frac{1}{L}\sum_{l=1}^{L}(E_s^l - R_s^l)^2. \tag{7}$$

As for the align head, we use the visual feature v_i^{clip} extracted from CLIP's native image encoder as the alignment target and optimize it via an alignment loss. This promotes semantic alignment between the decoder's visual output and the CLIP embedding space. Similar alignment strategies have been employed in ViLD (Gu et al. 2021) and OVTR, enabling models to better exploit CLIP's open-vocabulary capabilities by aligning visual features with its semantic space. The alignment loss is defined as follows:

$$\mathcal{L}_{\text{align}} = \frac{1}{K}\sum_{i=1}^{K}(\widetilde{v}_i - v_i^{\text{clip}})^2. \tag{8}$$

For bounding-box prediction, we adopt a combination of GIoU loss and L1 loss, defined as:

$$\mathcal{L}_{\text{box}} = \frac{1}{K}\sum_{i=1}^{K}\left(\lambda_g \mathcal{L}_g(\widehat{b}_i, b_i^{\text{gt}}) + \lambda_{\mathcal{L}_1}\mathcal{L}_1(\widehat{b}_i, b_i^{\text{gt}})\right), \tag{9}$$

where b_i^{gt} is the ground-truth box for the i-th object, and $\mathcal{L}_g(\cdot,\cdot)$, $\mathcal{L}_1(\cdot,\cdot)$ denote GIoU and L1 loss, respectively. Box matching is performed via the Hungarian algorithm, with loss weights λ_g and $\lambda_{\mathcal{L}_1}$ set to 2 and 1.

For classification, we apply Focal Loss to supervise the output predictions, which is defined as:

$$\mathcal{L}_{\text{cls}} = \text{FocalLoss}(\mathcal{Z}, \mathcal{Z}^{target}). \tag{10}$$

Combining the components described above, the overall loss function used for optimization is as follows:

$$\mathcal{L} = \lambda_1\mathcal{L}_{\text{cls}} + \lambda_2\mathcal{L}_{\text{box}} + \lambda_3\mathcal{L}_{\text{align}} + \lambda_4\mathcal{L}_{SEA}, \tag{11}$$

where λ_1, λ_2, λ_3, λ_4 are the weights for each loss component. Additionally, we apply auxiliary losses after each decoder layer to further optimize the decoder, using the same loss formulation as Eq. 11 except for $\mathcal{L}_{SEA}$

Sparse Detection for Inference. During inference, our method leverages the zero-shot tracking capability of SAM2 to perform temporal localization and identity association from an initial prompt, without requiring dense per-frame detection. To improve efficiency, we use a sparse detection strategy that detects only on keyframes at fixed intervals.

Tracking begins by applying UDM on the first frame to generate bounding-box prompts for SAM2. In subsequent frames, sparse detection is periodically applied to discover new objects. On each selected frame, predicted boxes are matched with existing SAM2 masks using IoU-based association. The matched detections inherit their corresponding IDs, while high-confidence unmatched detections are assigned new IDs and added as prompts to the tracking process through the prompt encoder. This sparse detection strategy reduces redundant computation while maintaining timely object discovery and consistent identity assignment.

Experiment

Experiment Setting

Datasets. We evaluate our method on TAO (Dave et al. 2020) for open-vocabulary tracking and KITTI (Geiger, Lenz, and Urtasun 2012) for cross-dataset zero-shot transfer. TAO contains 833 categories and follows the LVIS split, using base classes for training and novel ones for evaluation. KITTI focuses on autonomous driving scenarios with cars and pedestrians, and it is used to assess our model's generalization without any fine-tuning. Our model is trained on LVIS, which includes 1,203 categories (886 base, 317 novel), excluding novel classes during training.

Metrics. We use the TETA metric (Li et al. 2022) as a comprehensive measure of localization (LocA), association (AssA), and classification (ClsA) accuracy on TAO. For KITTI, we report standard MOT metrics including MOTA, IDF1, and ID switches (IDs) to assess zero-shot ability under real-world conditions.

Implementation Details. We use SAM2.1-baseplus as the foundation module and build UDM with 6 encoder and decoder layers. SEA distillation adopts a DINO-based ResNet-50 as the teacher. The model is trained on static LVIS images using Adam for 30 epochs with a batch size of 4. Loss weights λ_1–λ_4 are set to 2, 5, 2, and 2 to emphasize localization. During inference, the prompt interval is 5, the instance cap per frame is 45, α is 0.1, and T_{w} is 10, reducing redundant detection while preserving accuracy.

State-of-the-Art Comparison

To validate the effectiveness of our method, we conduct both qualitative and quantitative analyses on both base and novel categories of the TAO validation and test sets, and compare against state-of-the-art methods.

The open-vocabulary tracking results in Table 1 show that SAM2-OV consistently outperforms state-of-the-art methods on both the validation and test splits of the TAO dataset. On the validation set, it achieves the highest TETA scores on both base and novel categories, surpassing the best baseline by 7.0% and 21.9%, respectively. For ClsA on novel categories, SAM2-OV improves over OVTR by 37%, while maintaining strong LocA and AssA. On the test set, it further improves base-category performance compared to the validation set and outperforms SLAck by 5.7% on novel categories. These results highlight SAM2-OV's robustness and stability in diverse open-world tracking scenarios.

Closed-set Evaluation. We further compare the overall performance of our method under the closed-set setting of the TAO benchmark using the TETA metric. As shown in Table 2, SAM2-OV demonstrates clear advantages in both LocA and AssA, with improvements of 0.9% and 12.3% over SLAcK, respectively. Our method further achieves strong association accuracy, demonstrating effective open-vocabulary instance modeling. Moreover, SAM2-OV is practically efficient, since it tunes only the detection stage: compared with OVTR under the same hardware, it converges faster (197→96 h), requires less inference memory (23.5→6.1 GB), and runs slightly faster (2.8→3.2 FPS).

Method	Training Data		Novel				Base			
Validation set	LVIS	TAO	TETA	LocA	AssocA	ClsA	TETA	LocA	AssocA	ClsA
DeepSORT (ViLD) (Wojke et al. 2017)	✓	✓	21.1	46.4	14.7	2.3	26.9	47.1	15.8	17.7
QDTrack* (Fischer et al. 2023)	✓	✓	22.5	42.7	24.4	0.4	27.1	45.6	24.7	11.0
TETer* (Li et al. 2022)	✓	✓	25.7	45.9	31.1	0.2	30.3	47.4	31.6	12.1
OVTrack (Li et al. 2023)	✓	-	27.8	48.8	33.6	1.5	35.5	49.3	36.9	**20.2**
MASA (Li et al. 2024b)	✓	-	30.0	54.2	34.6	1.0	36.9	**55.1**	36.4	19.3
OVTR (Li et al. 2025b)	✓	-	31.4	54.4	34.5	5.4	36.6	52.2	37.6	20.1
SLAck (Li et al. 2024c)	✓	-	31.1	54.3	37.8	1.3	37.2	55.0	37.6	19.1
SAM2-OV (Ours)	✓	-	**37.9**	**57.6**	**48.5**	**7.4**	**39.8**	52.4	**51.6**	15.4
Test set	LVIS	TAO	TETA	LocA	AssocA	ClsA	TETA	LocA	AssocA	ClsA
DeepSORT (ViLD) (Wojke et al. 2017)	✓	✓	17.2	38.4	11.6	1.7	24.5	43.8	14.6	15.2
QDTrack* (Fischer et al. 2023)	✓	✓	20.2	39.7	20.9	0.2	25.8	43.2	23.5	10.6
TETer* (Li et al. 2022)	✓	✓	21.7	39.1	25.9	0.0	29.2	44.0	30.4	10.7
OVTrack (Li et al. 2023)	✓	-	24.1	41.8	28.7	1.8	32.6	45.6	35.4	**16.9**
OVTR (Li et al. 2025b)	✓	-	27.1	47.1	32.1	**2.1**	34.5	51.1	37.5	14.9
SLAck (Li et al. 2024c)	✓	-	27.1	49.1	30.0	2.0	34.7	52.5	35.6	16.1
SAM2-OV (Ours)	✓	-	**32.8**	**52.6**	**44.1**	1.7	**41.0**	**55.6**	**55.6**	11.8

Table 1: Comparison with state-of-the-art methods on the TAO validation and test sets. All methods use the ResNet-50 backbone except SAM2-OV, distilled from ResNet-50. $*$ indicates training on both base and novel categories. Best results are in bold.

Method	TETA	LocA	AssocA	ClsA
DeepSORT (Wojke et al. 2017)	26.0	48.4	17.5	12.1
QDTrack (Fischer et al. 2023)	30.0	50.5	27.4	12.1
TETer (Li et al. 2022)	33.3	51.6	35.0	13.2
OVTrack (Li et al. 2023)	34.7	49.3	36.7	18.1
MASA(R50)† (Li et al. 2024b)	34.1	52.1	35.7	15.0
OVTR (Li et al. 2025b)	35.7	52.3	36.3	**18.4**
SLAcK-T† (Li et al. 2024c)	35.5	52.2	38.9	15.6
SAM2-OV (Ours)	**39.6**	**53.1**	**51.3**	14.5

Table 2: Comparison of closed-set MOT methods using the TETA metric on the TAO validation set. † indicates detection results using TETer-T; SLAcK-T uses Swin-T as the backbone. Best results are in bold.

Base	SEA	MPAM	TETA	LocA	AssocA	ClsA	ClsA_n
			16.5	5.8	43.8	-	-
✓			38.5	54.6	48.3	12.5	2.4
✓	✓		39.5	53.1	51.2	14.3	7.5
✓		✓	**40.8**	**55.2**	**54.3**	13.0	2.1
✓	✓	✓	39.6	53.1	51.3	**14.5**	**7.5**

Table 3: Ablation studies on the TAO validation set evaluating SAM2-OV modules. Base uses UDM without SEA and MPAM. ClsA_n denotes the ClsA on novel categories, and "–" indicates a non-applicable metric.

Ablation Study

OV-MOT with SAM2 basline. We use SAM2 with grid prompts as a baseline setting, where prompts are uniformly distributed across the image to initiate tracking. As classification is not involved, ClsA is omitted. Table 3 shows that SAM2 maintains a reasonable association but struggles with accurate localization without detection guidance.

The effect of using different modules. The results in Table 3 show that introducing SEA notably improves semantic modeling, raising ClsA from 12.5% to 14.3% and ClsA_n from 2.4% to 7.5%, contributing to a 1.0% gain in TETA. However, it slightly reduces LocA when combined with MPAM. In contrast, using MPAM alone significantly improves LocA and AssocA, achieving the highest TETA of 40.8%, but slightly lowers ClsA_n by 0.3% due to SAM2's limited semantic capacity, making temporal cues insufficient for accurate classification. When SEA and MPAM are combined, the model achieves balanced performance, integrating semantic generalization and temporal robustness. To further analyze semantic enhancement, we examine the distillation weight λ_4 in SEA. Increasing λ_4 enhances semantic transfer, improving open-vocabulary classification (e.g., ClsA_n increases from 5.1 to 7.5 when λ_4 rises from 1 to 2). However, stronger distillation slightly weakens spatial and temporal stability (TETA decreases from 40.1 to 39.2 at $\lambda_4 = 3$). We adopt a moderate λ_4 to balance semantic generalization and localization robustness.

Analysis of detection hyperparameters. To evaluate the practicality of UDM, we conduct a sensitivity analysis on two critical hyperparameters: the maximum number of detectable objects per frame (30, 45, 60) and the interval between prompt frames (1, 5, 7). As shown in Table 4, setting the maximum detection count too low (30) misses important objects, reducing detection and tracking quality, while setting it too high (60) introduces redundant low-confidence predictions and degrades performance. Similarly, prompting

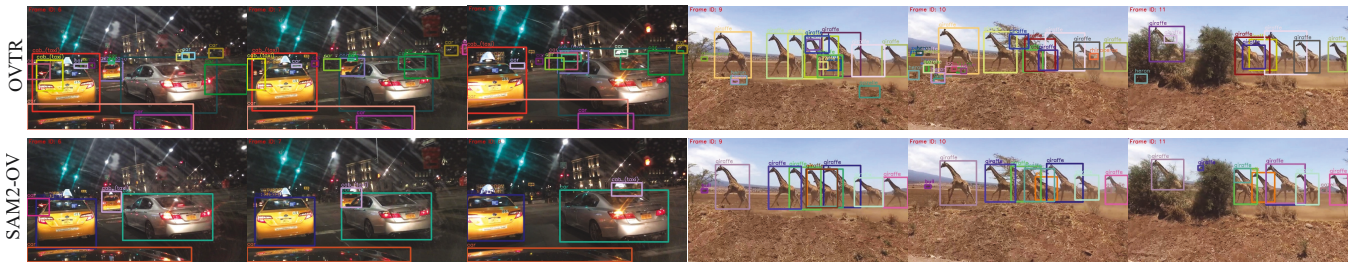

Figure 5: Qualitative results of SAM2-OV and OVTR methods. The examples include scenes with occlusion and viewpoint changes. The color of each box indicates the identity ID of the object.

Type	Setting	TETA	LocA	AssocA	ClsA
Max Number Per Frame	30	39.4	52.5	**51.9**	13.9
	45	**39.6**	**53.1**	51.3	**14.5**
	60	39.2	51.9	51.4	14.4
Detection Interval	1	**39.7**	52.2	**53.7**	13.3
	5	39.6	53.1	51.3	**14.5**
	7	39.2	**53.6**	51.9	12.0

Table 4: Analysis of the impact of the maximum number of detectable instances per frame and the detection interval on SAM2-OV performance.

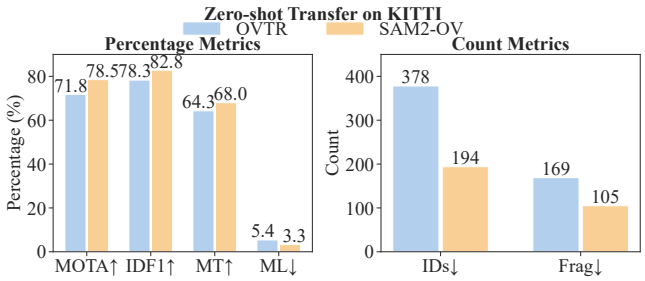

Figure 6: Zero-shot domain transfer performance on the KITTI dataset, evaluated using OVTR and SAM2-OV.

the model at every frame provides minimal accuracy gains yet significantly increases computational load due to redundancy, whereas prompting too infrequently (interval of 7 frames) delays discovering new objects and negatively affects tracking quality. Using 45 objects per frame and a 5-frame prompt interval achieves a good trade-off between accuracy and efficiency, supporting real-world deployment.

Analysis of motion modeling hyperparameters. To evaluate the impact of motion information on mask selection, we analyze the weighting coefficient in the Motion Prior Assistance Module, which balances the Kalman-predicted IoU and SAM2's original IoU. As shown in Table 5, a high weight (0.5) leads to over-reliance on motion cues, degrading semantic accuracy and reducing TETA to 38.9. Lowering the coefficient improves performance, with the best result (TETA = 39.6) achieved at 0.1, surpassing the baseline without motion modeling. These results demonstrate that a properly tuned motion prior improves mask selection robustness, especially under occlusion or rapid motion.

MPAM Weight	TETA	LocA	AssocA	ClsA
-	39.5	53.1	51.2	14.3
0.1	**39.6**	**53.1**	51.3	**14.5**
0.2	39.5	52.5	**52.1**	13.8
0.5	38.9	51.9	51.1	13.7

Table 5: Effect of the motion prior fusion weight in MPAM on SAM2-OV performance.

Zero-Shot Transfer on KITTI

We evaluate the zero-shot transfer performance of SAM2-OV on the KITTI dataset in Figure 6, comparing it with OVTR. Due to ambiguity in the definition of the Pedestrian class between KITTI and open-vocabulary settings, we report results only for the car category. As shown in the figure, SAM2-OV outperforms OVTR by 9.3% in MOTA and reduces ID switches by 48.7%, demonstrating robust and accurate tracking across domains, as well as strong zero-shot generalization in cross-dataset scenarios.

Qualitative Results

To clearly demonstrate the advantages of our method, we visualize results from OVTR and our SAM2-OV. As shown in Figure 5, compared to OVTR, SAM2-OV produces fewer false positives on background regions and achieves more robust and stable tracking under occlusion and viewpoint changes, showing superior identity preservation.

Conclusion

In this paper, we present SAM2-OV, a novel open-vocabulary multi-object tracking method built upon a detection-only tuning paradigm. By training solely on static images without using association-specific supervision, SAM2-OV simplifies the training process while maintaining strong tracking performance. The method leverages SAM2's cross-frame matching capabilities for robust identity association and integrates motion modeling and semantic enhancement to improve stability and generalization. SAM2-OV enables efficient inference through sparse prompting and shows strong zero-shot transferability, which demonstrates its versatility and robustness in real-world applications.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants U25A20531, U21A20514 and 62502402, and in part by the Major Science and Technology Plan Project on the Future Industry Fields of Xiamen City under Grant 3502Z20241027.

References

Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; and Upcroft, B. 2016. Simple Online and Realtime Tracking. In *Proceedings of IEEE International Conference on Image Processing*, 3464–3468.

Cao, J.; Pang, J.; Weng, X.; Khirodkar, R.; and Kitani, K. 2023. Observation-centric sort: Rethinking Sort for Robust Multi-Object Tracking. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9686–9696.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In *Proceedings of European Conference on Computer Vision*, 213–229.

Dave, A.; Khurana, T.; Tokmakov, P.; Schmid, C.; and Ramanan, D. 2020. TAO: A Large-Scale Benchmark for Tracking Any Object. In *Proceedings of European Conference on Computer Vision*, 436–454.

Du, Y.; Wei, F.; Zhang, Z.; Shi, M.; Gao, Y.; and Li, G. 2022. Learning to Prompt for Open-Vocabulary Object Detection with Vision-Language Model. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14084–14093.

Fang, R.; Pang, G.; and Bai, X. 2024. Simple Image-level Classification Improves Open-Vocabulary Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1716–1725.

Fischer, T.; Huang, T. E.; Pang, J.; Qiu, L.; Chen, H.; Darrell, T.; and Yu, F. 2023. QDTrack: Quasi-Dense Similarity Learning for Appearance-Only Multiple Object Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12): 15380–15393.

Gao, R.; and Wang, L. 2023. MeMOTR: Long-Term Memory-Augmented Transformer for Multi-Object Tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9901–9910.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3354–3361.

Gu, X.; Lin, T.-Y.; Kuo, W.; and Cui, Y. 2021. Open-Vocabulary Object Detection via Vision and Language Knowledge Distillation. *arXiv preprint arXiv:2104.13921*.

Jiang, J.; Wang, Z.; Zhao, M.; Li, Y.; and Jiang, D. 2025. SAM2MOT: A Novel Paradigm of Multi-Object Tracking by Segmentation. *arXiv preprint arXiv:2504.04519*.

Li, G.; Jian, Y.; Jian, Y.; Yan, Y.; Yan, Y.; Wang, H.; and Wang, H. 2024a. GLATrack: Global and Local Awareness for Open-Vocabulary Multiple Object Tracking. In *Proceedings of ACM International Conference on Multimedia*, 2457–2466.

Li, G.; Zhuang, S.; Jian, Y.; Yan, Y.; and Wang, H. 2025a. Language Decoupling with Fine-grained Knowledge Guidance for Referring Multi-object Tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23626–23635.

Li, J.; Yu, E.; Chen, S.; and Tao, W. 2025b. OVTR: End-to-End Open-Vocabulary Multiple Object Tracking with Transformer. *arXiv preprint arXiv:2503.10616*.

Li, S.; Danelljan, M.; Ding, H.; Huang, T. E.; and Yu, F. 2022. Tracking Every Thing in The Wild. In *Proceedings of European Conference on Computer Vision*, 498–515.

Li, S.; Fischer, T.; Ke, L.; Ding, H.; Danelljan, M.; and Yu, F. 2023. Ovtrack: Open-Vocabulary Multiple Object Tracking. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5567–5577.

Li, S.; Ke, L.; Danelljan, M.; Piccinelli, L.; Segu, M.; Van Gool, L.; and Yu, F. 2024b. Matching Anything by Segmenting Anything. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18963–18973.

Li, S.; Ke, L.; Yang, Y.-H.; Piccinelli, L.; Segu, M.; Danelljan, M.; and Gool, L. V. 2024c. Slack: Semantic,Location, and Appearance Aware Open-vocabulary Tracking. In *Proceedings of European Conference on Computer Vision*, 1–18.

Lin, C.; Sun, P.; Jiang, Y.; Luo, P.; Qu, L.; Haffari, G.; Yuan, Z.; and Cai, J. 2022. Learning Object-Language Alignments for Open-Vocabulary Object Detection. *arXiv preprint arXiv:2211.14843*.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2024. Grounding Dino: Marrying Dino with Grounded Pre-Training for Open-Set object Detection. In *Proceedings of European Conference on Computer Vision*, 38–55.

Liu, Y.; Zulfikar, I. E.; Luiten, J.; Dave, A.; Ramanan, D.; Leibe, B.; Ošep, A.; and Leal-Taixé, L. 2022. Opening Up Open World Tracking. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19045–19055.

Luo, R.; Song, Z.; Ma, L.; Wei, J.; Yang, W.; and Yang, M. 2024. Diffusiontrack: Diffusion Model for Multi-Object Tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3991–3999.

Ma, S.; Qian, D.; Ye, K.; and Zhang, S. 2025. Cake: Category Aware Knowledge Extraction for Open-Vocabulary Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 5982–5990.

Ma, Y.; Tang, Y.; Yang, W.; Zhang, T.; Zhang, J.; and Kang, M. 2024. Unifying Visual and Vision-Language Tracking via Contrastive Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 4107–4116.

Meinhardt, T.; Kirillov, A.; Leal-Taixe, L.; and Feichtenhofer, C. 2022. Trackformer: Multi-Object Tracking with

Transformers. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8844–8854.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. 2024. Sam 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714*.

Song, Z.; Luo, R.; Ma, L.; Tang, Y.; Chen, Y.-P. P.; Yu, J.; and Yang, W. 2025. Temporal Coherent Object Flow for Multi-Object Tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 6978–6986.

Tang, Z.; Xu, T.; Wu, X.; Zhu, X.-F.; and Kittler, J. 2024. Generative-Based Fusion Mechanism for Multi-Modal Tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 5189–5197.

Videnovic, J.; Lukezic, A.; and Kristan, M. 2025. A Distractor-Aware Memory for Visual Object Tracking with SAM2. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24255–24264.

Vo, D.-K.; Nguyen, V.-L.; Tran, M.-T.; and Le, T.-N. 2025. SAMURAI: Shape-Aware Multimodal Retrieval for 3D Object Identification. *arXiv preprint arXiv:2506.21056*.

Wojke, N.; Bewley, A.; Paulus, D.; and test, t. 2017. Simple Online and Realtime Tracking with A Deep Association Metric. In *Proceedings of IEEE International Conference on Image Processing*, 3645–3649.

Wu, H.; Han, W.; Wen, C.; Li, X.; and Wang, C. 2021. 3D Multi-Object Tracking in Point Clouds Based on Prediction Confidence-guided data Association. *IEEE Transactions on Intelligent Transportation Systems*, 23(6): 5668–5677.

Wu, Q.; Wan, J.; and Chan, A. B. 2021. Progressive Unsupervised Learning for Visual Object Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2993–3002.

Yi, K.; Luo, K.; Luo, X.; Huang, J.; Wu, H.; Hu, R.; and Hao, W. 2024. Ucmctrack: Multi-object Tracking with Uniform Camera Motion Compensation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 6702–6710.

Zeng, R.; Huang, Y.; and Pei, S. 2025. TGFormer: Transformer with Track Query Group for Multi-Object Tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 9824–9832.

Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, F.; Yuan, Z.; Luo, P.; Liu, W.; and Wang, X. 2022. Bytetrack: Multi-Object Tracking by Associating Every Detection Box. In *Proceedings of European Conference on Computer Vision*, 1–21.

Zhang, Y.; Wang, C.; Wang, X.; Zeng, W.; and Liu, W. 2021. Fairmot: On the Fairness of Detection and Re-identification in Multiple Object Tracking. *International Journal of Computer Vision*, 129(11): 3069–3087.

Zhou, X.; Girdhar, R.; Joulin, A.; Krähenbühl, P.; and Misra, I. 2022. Detecting Twenty-Thousand Classes Using Image-Level Supervision. In *Proceedings of European Conference on Computer Vision*, 350–368.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv preprint arXiv:2010.04159*.

Zhuang, S.; Li, G.; Wu, Q.; Lu, Y.; Hu, H.-M.; and Wang, H. 2025. CGATracker: Correlation-Aware Graph Alignment for Referring Multi-Object Tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(11): 11337–11349.