

# Language Decoupling with Fine-grained Knowledge Guidance for Referring Multi-object Tracking

Guangyao Li<sup>1,2</sup>, Siping Zhuang<sup>1,2</sup>, Yajun Jian<sup>1,2</sup>, Yan Yan<sup>1,2</sup>, Hanzi Wang<sup>1,2\*</sup>

<sup>1</sup> Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, 361005, P.R. China

<sup>2</sup> Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, 361005, P.R. China

liguangyao@stu.xmu.edu.cn, seepyzhuang@stu.xmu.edu.cn, yajunjian@stu.xmu.edu.cn

yanyan@xmu.edu.cn, hanzi.wang@xmu.edu.cn

## Abstract

Referring multi-object tracking (RMOT) aims to detect and track specific objects based on natural language expressions. Previous methods typically rely on sentence-level vision-language alignment, often failing to exploit fine-grained linguistic cues that are crucial for distinguishing objects with similar characteristics. Notably, these cues play distinct roles at different tracking stages and should be leveraged accordingly to provide more explicit guidance. In this work, we propose DKGTrack, a novel RMOT method that enhances language comprehension for precise object tracking by decoupling language expressions into localized descriptions and motion states. To improve the accuracy of language-guided object identification, we introduce a Static Semantic Enhancement (SSE) module, which enhances region-level vision-language alignment through hierarchical cross-modal feature interaction, providing more discriminative object representations for tracking. Furthermore, we propose a Motion Perception Alignment (MPA) module that explicitly aligns object queries with motion descriptions, enabling accurate object trajectory prediction across frames. Experimental results on multiple RMOT benchmarks demonstrate the effectiveness of our method, which achieves competitive performance in challenging tracking scenarios. The code is available at <https://github.com/acyddl/DKGTrack>.

## 1. Introduction

Referring multi-object tracking (RMOT) aims to detect and track objects of interest in video sequences guided by natural language expressions. This emerging task bridges com-

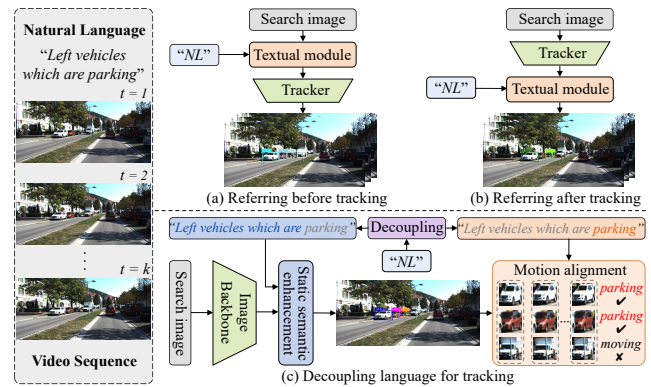


Figure 1. Comparison of different pipelines using natural language as referring guidance. Previous methods rely solely on video-level referring expression comprehension, leading to incorrect associations in complex scenarios. In contrast, our method decouples language and uses fine-grained knowledge as referring guidance, giving more accurate tracking performance.

puter vision and natural language processing, representing a critical vision-language problem in multi-modal information interaction. RMOT has gained significant research attention due to its potential in real-world applications, such as video editing [14] and autonomous driving [49]. Although RMOT shares historical roots with traditional multi-object tracking, it distinguishes itself by leveraging natural language expressions to track referred objects. Notably, recent datasets like Refer-KITTI-V2 [48] have highlighted the importance of semantic diversity in this field, emphasizing the necessity to comprehend language expressions close to real-world scenarios.

Existing RMOT methods typically transform natural language expressions into a single sentence embedding as the reference. For instance, TransRMOT [40] and TempRMOT [48] employ a cross-modal encoder to fuse whole sentence

\*Corresponding author.

embeddings with image-level features for language comprehension. iKUN [7] adopts a two-stage paradigm, which first generates object tracklets and then selects the most relevant ones based on the given language expressions. However, these methods rely on coarse similarity measurements between targets and sentence embeddings, failing to exploit fine-grained linguistic cues. This limitation makes it challenging to distinguish objects with highly similar sentence embeddings. Moreover, objects in video sequences exhibit diverse motion states and localized descriptions. Since image-level features are insufficient to capture motion states, and localized descriptions may overshadow the motion cues, simply combining the two may result in ambiguous target identification. As shown in Fig. 1(a), when the white vehicle is moving on the left, the tracker incorrectly identifies it as parking. Furthermore, using the entire sentence to match the existing trajectories makes it difficult for the model to align language expressions with specific targets. As shown in Fig. 1(b), this misalignment confuses the tracker during the matching process, resulting in the missing of the red vehicle parking on the left.

In this paper, we propose DKGTrack, a novel referring multi-object tracking method that decouples language and leverages fine-grained knowledge as referring guidance to address the above limitations. By precisely integrating semantic information into the tracker, DKGTrack enhances comprehension of natural language and enables accurate object tracking. Specifically, DKGTrack uses a pre-trained language model to extract motion states and localized descriptions from language expressions. Localized descriptions are used to detect potential candidates based on static visual features within each frame, while motion states distinguish candidates that conform to the described motion. This enables localized descriptions and motion states to complement each other, improving the understanding of referring expressions and video content. To precisely capture and align region-level features with localized descriptions, we propose a Static Semantic Enhancement (SSE) module, which enables interactions between sentence embedding and initialized queries to generate appropriate queries for tracking. Next, we introduce a Recurrent Text Guidance (RTG) strategy to enhance alignment between the refined image features generated by the Transformer encoder and the localized descriptions. Unlike previous image-level feature fusion methods, RTG complements image features with localized descriptions to obtain more comprehensive representations. These enhanced features guide the learning of local features, where target queries interact with refined image features to generate discriminative target representations. Finally, the correlation similarity between region-level features and individual words in the sentence is calculated to minimize redundant background interference.

Moreover, to alleviate the challenge of aligning track-

lets with object motion states across the temporal domain, we propose a Motion Perception Alignment (MPA) module. The MPA module leverages motion descriptions from language expressions to assist object queries in comprehending temporal information. By aligning motion descriptions with object queries, MPA bridges the gap between visual and language modalities, enabling a coherent understanding of object trajectories. As a result, DKGTrack significantly improves the matching accuracy of object tracklets, ensuring stable tracking in complex environments. Furthermore, by integrating motion information with object queries generated by the Transformer decoder, DKGTrack enhances motion awareness, improving detection and tracking accuracy in the current frame while selecting matching object tracklets. The main contributions are summarized as follows:

- We propose DKGTrack, a novel referring multi-object tracking method that decouples language and leverages fine-grained knowledge as guidance, achieving exceptional performance in tracking specific objects in videos.
- We introduce a static semantic enhancement module to improve the model’s ability to identify referred regions at the image level and a recurrent text guidance strategy to further capture local expressions.
- We propose a motion perception alignment module, which enhances temporal awareness of objects. This enables more accurate object association based on motion states obtained from language expressions.
- Extensive experiments on multiple benchmarks show that DKGTrack outperforms existing methods and achieves significant performance gains in challenging scenarios.

## 2. Related work

**Multi-Object Tracking.** Multi-object tracking [13, 17, 31, 39, 43, 44] aims to simultaneously track multiple objects in complex scenarios. Most existing methods [3, 10, 19, 29, 30, 35] adopt the tracking-by-detection paradigm [8, 18], which first detects objects in individual frames and then associates them based on their identities. For example, ByteTrack [46] links nearly all detected bounding boxes rather than restricting associations to high-confidence detections, improving the recall performance in crowded scenes. OC-SORT [1] enhances traditional Kalman filter-based tracking by incorporating object observations to correct error accumulations during occlusions, improving robustness against non-linear motion. DiffMOT [26] introduces a decoupled diffusion-based motion predictor to model diverse non-linear object motions. However, these methods rely on separate detection and tracking phases, which often lead to high computational costs and complexity. To address these limitations, end-to-end methods integrate object detection and tracking into a unified framework, enabling joint optimization and improving tracking efficiency. For instance, PermaTrack [36] incorporates a spatial-temporal recurrent

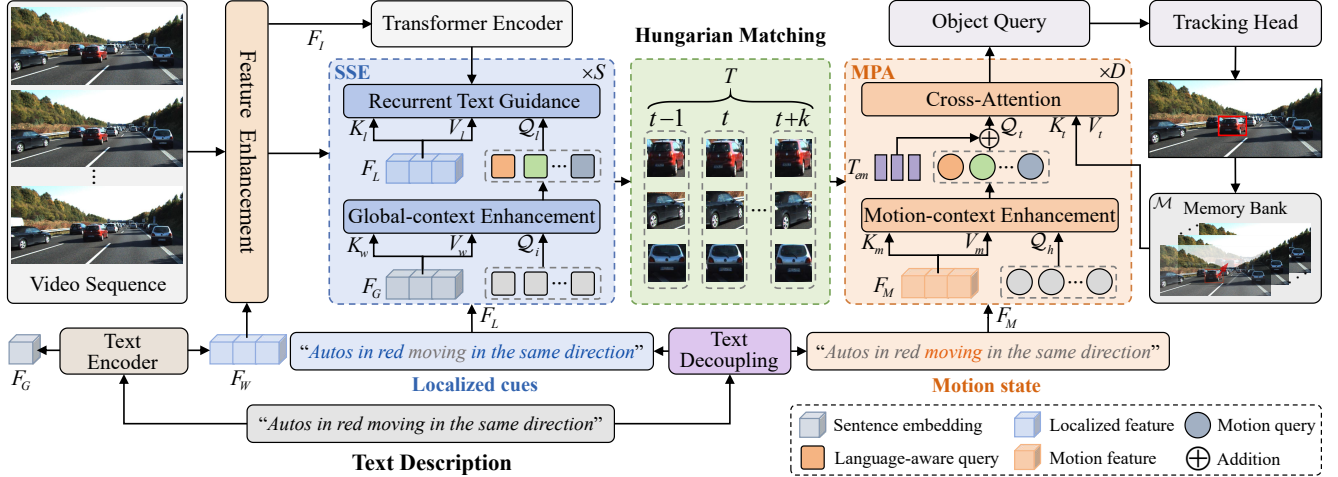


Figure 2. The overall architecture of the proposed DKGTrack. First, the image backbone and language model extract discrete visual embeddings and text embeddings for a given image and corresponding text prompts, respectively. Second, the motion states and localized descriptions are extracted from text embeddings via a pre-trained language model. Then, the object queries are refined through global-context interaction, and the recurrent text guidance strategy is used to enhance alignment between image features and language descriptions. Finally, the motion perception alignment module aligns motion descriptions with object queries, ensuring coherent trajectory prediction.

memory module to reason about object permanence, enabling robust tracking even under full occlusion. Similarly, Memotr [9] enhances object association by stabilizing and distinguishing track embeddings through long-term memory infusion with a customized memory-attention layer. Although these methods improve tracking robustness, they fundamentally operate under traditional paradigms that indiscriminately track all detected objects. In contrast, the proposed DKGTrack introduces explicit cross-modal interaction mechanisms to achieve fine-grained vision-language alignment, enabling the precise tracking of referred objects in challenging scenarios.

**Referring Understanding.** Referring understanding [6, 16, 42, 47, 50] has emerged as a significant research task in recent years, aiming to identify specific objects using natural language or audio descriptions. For image-based methods, APE [33] introduces a universal visual perception model that simultaneously aligns all objects in an image with text prompts for object detection and instance segmentation. PPT [4] proposes a point generator and a multi-source curriculum learning strategy to enhance segmentation precision by integrating the CLIP’s text-image alignment and SAM’s mask generation capabilities. GroundingREC [5] integrates the local subject with low-layer features and global context with high-layer features, enabling fine-grained object counting within the same class based on language descriptions. ReferDiffusion [23] employs a multi-stage training strategy to improve the identification accuracy of specific objects using text prompts. For video-based methods, OnlineRefer [41] enhances video-based referring description comprehension by leveraging dense feature aggregation and cross-modal fusion. TransRMOT [40]

and TempRMOT [48] integrate textual modules into existing trackers to track objects of interest with language guidance. UVLTrack [27] seamlessly handles the target references with a modality-adaptive design for robust tracking. QueryNLT [32] integrates temporal visual templates and linguistic expressions through prompt modulation, ensuring spatial-temporal consistency and accurate target localization. However, these methods struggle to comprehend fine-grained textual cues and leverage precise language expressions to localize referred objects and infer their motion states. To address these limitations, our method decouples language information into localized descriptions and motion states to better understand the referring language.

## 3. Method

### 3.1. Overview

We propose DKGTrack, which leverages fine-grained knowledge derived from language descriptions to enhance tracking performance through two key components: the Static Semantic Enhancement (SSE) module, which improves image-level localization, and the Motion Perception Alignment (MPA) module, which incorporates motion information into the tracking process. Specifically, given a video sequence  $\mathcal{V}$  consisting of  $T$  frames and the corresponding referring language expression  $\mathcal{E}$  consisting of  $L$  words, we first extract image features  $F_S$  using ResNet50 [11] and linguistic features using the pre-trained RoBERTa model [24], which generates a sentence embedding  $F_G$  and word-level features  $F_W$ . Next, the  $F_W$  are decoupled into motion states  $F_M$  and localized descriptions  $F_L$  using a binary mask  $M_v$ , which is generated by analyzing the lin-

guistic properties of each word. The object queries  $Q_i$  are initialized by the sentence embedding  $F_G$ , while image features  $F_S$  undergo cross-modal interaction with word-level features  $F_W$  to generate modality-enhanced visual features  $F_I$ . Subsequently, the RTG strategy progressively refines  $Q_i$  using localized descriptions  $F_L$  to capture spatial information about candidates. Finally, the MPA module embeds motion information into queries. These queries are then used to construct a memory bank  $\mathcal{M}$ , which stores query embeddings from previous frames and updates object queries in the current frame to improve tracking robustness.

### 3.2. Text Decoupling

Existing referring multi-object tracking methods typically encode sentences into single-sentence embeddings. However, these embeddings lack fine-grained understanding and have difficulty in handling complex sentence structures. To address this limitation, we decouple the whole sentence  $\mathcal{E}$  into localized descriptions  $F_L$  and motion states  $F_M$ , enabling a more precise understanding of the sentence components relevant to object localization and motion characteristics. This process can be formalized as follows:

$$F_M = M_v \odot \mathcal{E}, \quad F_L = (1 - M_v) \odot \mathcal{E}, \quad (1)$$

where  $M_v \in \mathbb{R}^{L \times 1}$  is a binary mask automatically generated based on part-of-speech tags of the language  $\mathcal{E}$ . Each element  $m_i \in M_v$  is 1 for motion-relevant embeddings and 0 for other embeddings.  $\odot$  denotes the element-wise multiplication.  $F_L$  captures static attributes such as object category, color and relative position, while  $F_M$  contains motion states like object dynamics. These embeddings are processed in different branches:  $F_L$  refines object queries for precise localization, while  $F_M$  is integrated into the MPA module to improve the temporal consistency of targets.

### 3.3. Static Semantic Enhancement

We propose the SSE module to align image-level features with language representations to extract the most relevant visual cues, which effectively alleviates the complicating object localization caused by the interference of motion descriptions. Given a set of queries  $Q_i$  and refined image-level features  $F'_I$  generated by the Transformer encoder. In order to clarify the targets for tracking, we integrate semantic information from the sentence embedding  $F_G$  into the  $Q_i$  through the global context enhancement process to generate language-aware object queries:

$$Q_i = Q_i + \sigma \text{Softmax} \left( \frac{Q_i F_G^T}{\sqrt{d}} \right) F_G, \quad (2)$$

where  $Q_i$  is used as the input to the first layer of the decoder.  $\sigma$  is a temperature parameter controlling the degree of semantic enhancement, and  $d$  is a scaling factor to stabilize attention scores.

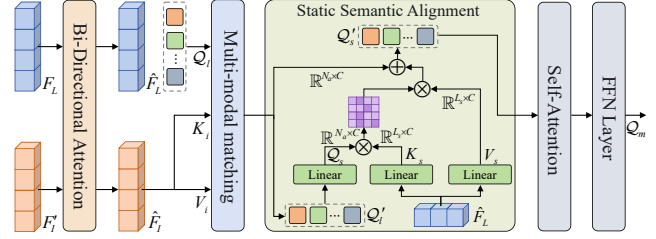


Figure 3. An illustration of the recurrent text guidance mechanism, which iteratively refines object queries using a bidirectional attention mechanism to fuse image features  $F'_I$  and word-level features  $F_L$  followed by multi-modal matching to focus on semantically relevant regions. Then the self-attention mechanism and a feed-forward network are used to refine the queries, ensuring robust spatial and semantic alignment for accurate tracking.

To further refine the cross-modal alignment, we introduce a recurrent text guidance strategy that iteratively updates object queries using both visual and language features. As shown in Fig. 3, we first use a bidirectional attention mechanism  $BiAttn(\cdot, \cdot)$  to complement the image features  $F'_I$  and the word-level features  $F_L$  as follows:

$$\hat{F}_I, \hat{F}_L = BiAttn(F'_I, F_L), \quad (3)$$

where  $BiAttn(\cdot, \cdot)$  computes the attention weights between the two modalities, enhancing alignment between visual features and localized descriptions.  $\hat{F}_I$  represents the image features enriched with linguistic information, while  $\hat{F}_L$  denotes the word-level features refined by using the visual context information. Next, the object queries  $Q_i$  are updated by interacting with the enriched visual features  $\hat{F}_I$  using a deformable attention mechanism as follows:

$$Q'_i = DeformAttn(\hat{F}_I, Q_i), \quad (4)$$

where  $Q'_i$  represents queries aligned with visual information, and  $DeformAttn(\cdot, \cdot)$  denotes the combination of the self-attention mechanism and the cross-attention mechanism. This ensures the queries focus on semantically relevant regions, improving the localization accuracy of the referred objects. Subsequently, we refine the queries using the localized descriptions  $\hat{F}_L$ :

$$Q'_s = Q'_i + \text{Softmax} \left( \frac{f(Q'_i, \hat{F}_L)}{\sqrt{d}} \right) \hat{F}_L, \quad (5)$$

where  $Q'_s$  represents the language-enhanced object queries, and  $f(\cdot, \cdot)$  denotes the similarity measure. This step effectively bridges the gap between vision and localized description, ensuring semantic accuracy in object queries. Finally, the self-attention mechanism is applied to model inter-object relationships, followed by a feedforward network to refine the object representations:

$$Q_m = FFN(\text{SelfAttn}(Q'_s)), \quad (6)$$

where  $Q_m$  represents the object queries generated by the static semantic enhancement module. Such a way ensures effective cross-modal feature fusion, query enhancement and multi-object localization, improving the robustness of referring multi-object tracking.

### 3.4. Motion Perception Alignment

A significant challenge in referring multi-object tracking lies in accurately identifying and aligning object motions across multiple frames. Previous methods [37, 48] simply utilize the coarse temporal encoding to enhance the perception of temporal dynamics. However, those methods only improve the temporal perception of all candidates without paying attention to the queried objects. Consequently, they may fail to effectively align the dynamic changes of targets with the corresponding language expressions. Notably, language descriptions inherently provide motion cues across different time spans, such as the short-term or long-term motion of the targets. To address this limitation, we propose a motion perception alignment module, which aligns language expressions with target trajectories by enhancing the motion properties of object queries. Specifically, we first use the Hungarian matching algorithm [15]  $Matcher(\cdot, \cdot)$  to associate the objects  $Q_m$  detected in the current frame with existing tracklets:

$$\begin{cases} Q_h^t = Matcher(\tilde{Q}_m^{t-1}, Q_m^t), & t \in [2, T] \\ Q_h^t = Q_m^t, & t = 1 \end{cases}, \quad (7)$$

where  $Q_h$  denotes object queries related to the localized descriptions. In this way, we obtain the object trajectories  $\mathcal{T}$  in the current frame. Subsequently, the motion states  $F_M$  is used to update object queries  $Q_h$ :

$$Q_h' = Q_h + \beta \text{Softmax} \left( \frac{f(Q_h, F_M^T)}{\sqrt{d}} \right) F_M, \quad (8)$$

where  $Q_h'$  is the motion-enhanced object queries, and  $\beta$  is a temperature parameter controlling the degree of motion state enhancement. Next, we use the temporal positional encoding to further enhance temporal awareness of the object representation. We construct a normalized temporal sequence as follows:

$$t_s = \frac{2\pi k}{T-1}, \quad k \in \{0, 1, \dots, T-1\} \quad (9)$$

where  $T$  is the total number of frames, and  $k$  represents the index of the current frame in the sequence. The temporal positional embedding is then computed as follows:

$$T_{em} = \begin{cases} \sin\left(\frac{t_s}{10000^{2\lfloor k/2 \rfloor / d}}\right), & k \bmod 2 = 0 \\ \cos\left(\frac{t_s}{10000^{2\lfloor k/2 \rfloor / d}}\right), & k \bmod 2 \neq 0 \end{cases}, \quad (10)$$

where  $d$  is the embedding dimension,  $\lfloor \cdot \rfloor$  denotes the floor function. The  $\sin(\cdot)$  and  $\cos(\cdot)$  functions are used to encode

the temporal information in a way that preserves both high-frequency and low-frequency patterns, enabling the model to capture temporal dynamics. Next, The motion-enhanced object queries  $Q_t$  are obtained as follows:

$$Q_t = Q_h' + T_{em}, \quad (11)$$

where the object queries  $Q_t$  are subsequently refined by a motion-state aware memory bank  $\mathcal{M}$  through a cross-attention mechanism, thereby enhancing the model's ability to capture temporal dependencies and understand complex temporal dynamics. The memory bank  $\mathcal{M}$  captures rich contextual information and contains object variations across previous frames. As newly detected objects are added, the oldest instances in the  $\mathcal{M}$  are progressively removed to maintain memory efficiency. By jointly modeling motion states and temporal embeddings, our MPA module effectively aligns language expressions with object trajectories, facilitating precise referring multi-object tracking.

### 3.5. Training Loss

The training loss of DKGTrack consists of localization loss, classification loss and referring loss. For precise object localization, we employ a combination of L1 loss and generalized IoU (GIoU) loss for bounding box regression:

$$\mathcal{L}_{loc} = \frac{1}{S} \sum_{i=1}^S (1 - \text{GIoU}(\hat{b}_i, b_i)) + \frac{1}{S} \sum_{i=1}^S \|\hat{b}_i - b_i\|_1, \quad (12)$$

where  $\hat{b}_i$  and  $b_i$  are the predicted and ground-truth bounding boxes, respectively.  $S$  is the total number of matched objects.  $\text{GIoU}(\cdot, \cdot)$  measures the intersection over the union of the predicted and ground-truth bounding boxes. This combination of losses ensures smooth penalization for misalignment in object localization. For classification, we use the focal loss [22] to handle class imbalance and emphasize hard-to-detect objects, improving robustness in cluttered or occluded scenes. In implementation, the focal loss is calculated as follows:

$$\mathcal{L}_{cls} = -\frac{1}{S} \sum_{i=1}^S \alpha_t (1 - p_t)^\gamma \log p_i^{y_i} (1 - p_i)^{(1-y_i)}, \quad (13)$$

where  $\alpha_t$  is a weighting factor that balances the contributions of positive and negative samples.  $\gamma$  is the focusing parameter that adjusts the relative weight of hard examples.  $y_i$  denotes the ground-truth labels, and  $p_i$  is the predicted class probability of the  $i^{\text{th}}$  object.  $p_t$  denotes the predicted confidence for a sample. Similar to the classification loss, we employ the focal loss as referring loss to improve the cross-modal correspondence between visual targets and language expressions. The overall loss is defined as follows:

$$\mathcal{L}_{loss} = \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{loc} + \lambda_3 \mathcal{L}_{ref}, \quad (14)$$

where  $\lambda_1$ ,  $\lambda_2$  and  $\lambda_3$  serve as weighting factors for  $\mathcal{L}_{cls}$ ,  $\mathcal{L}_{loc}$  and  $\mathcal{L}_{ref}$ , respectively.

Method	Detector	E	HOTA	DetA	AssA	DetRe	DetPr	AssRe	AssPr	LocA
FairMOT [45]	DLA-34	✗	23.46	14.84	40.15	17.40	43.58	53.35	73.15	74.77
DeepSORT [38]	DLA-34	✗	25.59	19.76	34.31	26.38	36.93	39.55	61.05	71.34
ByteTrack [46]	DLA-34	✗	22.49	13.17	40.62	16.13	36.61	46.09	73.39	73.90
CSTrack [2]	YOLOv8	✗	27.91	20.65	39.10	33.76	32.61	43.12	71.82	79.51
TransTrack [34]	DeformableDETR	✗	32.77	23.31	45.71	32.33	42.23	49.99	78.74	79.48
TrackFormer [28]	DeformableDETR	✗	33.26	25.44	45.87	35.21	42.19	50.26	78.92	79.63
iKUN [7]	DeformableDETR	✗	48.84	35.74	<b>66.80</b>	51.97	52.25	<b>72.95</b>	87.09	-
EchoTrack [20]	DeformableDETR	✓	39.47	31.19	51.56	42.65	48.86	56.68	81.21	79.93
DeepRMOT [12]	DeformableDETR	✓	39.55	30.12	53.23	41.91	47.47	58.47	82.16	80.49
TransRMOT [40]	DeformableDETR	✓	38.06	29.28	50.83	40.20	47.36	55.43	81.36	79.79
TransRMOT* [40]	DeformableDETR	✓	45.65	36.15	57.86	54.58	50.65	61.15	<b>89.96</b>	90.33
TempRMOT <sup>‡</sup> [48]	DeformableDETR	✓	50.31	38.48	65.93	52.23	57.88	70.98	87.76	90.40
DKGTrack (Ours)	DeformableDETR	✓	<b>52.08</b>	<b>41.10</b>	66.04	<b>57.57</b>	<b>58.36</b>	71.13	87.98	<b>90.54</b>

Table 1. Comparison with state-of-the-art MOT methods on the Refer-KITTI dataset. \* denotes the results after the frame correction operation. E means the End-to-end methods. ‡ indicates the results reproduced using official code. The best results are highlighted in bold.

## 4. Experiments

### 4.1. Datasets and Evaluation Protocol

**Datasets.** We evaluate our method on two datasets: Refer-KITTI [40] and its extended version Refer-KITTI-V2 [48]. Refer-KITTI is collected from the original KITTI dataset by removing three video sequences and adding manually annotated referring expressions. It contains 15 training videos with 660 unique descriptions and 3 test videos with 158 annotations. Refer-KITTI-V2 retains the full KITTI dataset and enriches it with LLM-generated referring annotations, including 17 training videos and 4 test videos.

**Metrics.** We use HOTA [25] as the primary evaluation metric, which decomposes tracking performance into Detection Accuracy (DetA) and Association Accuracy (AssA). DetA measures the spatial alignment between predictions and the ground truth using true positives, false negatives and false positives under an intersection-over-union (IoU) threshold. AssA evaluates temporal consistency by assessing correct, missed and erroneous identity associations.

### 4.2. Implementation Details

We use ResNet50 [11] as the visual backbone and RoBERTa [24] as the text encoder. The video frames are fed into the visual backbone to extract multi-scale features, which are projected into a shared embedding space with language features through linear layers for cross-modal feature extraction. The lengths of the memory bank  $\mathcal{M}$  for Refer-KITTI and Refer-KITTI-V2 are set to 4 and 5, respectively.

**Training.** We initialize the model parameters using Deformable DETR [51] pre-trained on the COCO dataset [21], while freezing the weights of RoBERTa. We use the AdamW optimizer with stepwise learning rates, which are

initially set to  $1 \times 10^{-5}$  for the visual backbone and the output heads, and  $1 \times 10^{-4}$  for the other network parameters. The training process runs with a batch size of 1 for 100 epochs on 4 NVIDIA RTX 3090 GPUs. The learning rates are reduced by a factor of 10 at the 40<sup>th</sup> epoch. The temperature parameters  $\alpha$  and  $\beta$  are set to 0.5. The loss weights are configured as  $\lambda_1 = 5$ ,  $\lambda_2 = 2$  and  $\lambda_3 = 2$ .

**Inference.** During inference, our model processes video sequences of arbitrary length along with referring expressions. For each frame, objects with the classification score higher than 0.6 are treated as foreground candidates. Tracking objects are considered lost if their classification scores remain below 0.4 for five consecutive frames.

### 4.3. Comparisons with State-of-the-art Methods

**Refer-KITTI.** The proposed DKGTrack demonstrates superior performance on the Refer-KITTI dataset, as shown in Tab. 1. It achieves the HOTA of 52.08% and the DetA of 41.10%, demonstrating its exceptional ability to accurately detect and track objects based on language descriptions. Although iKUN [7] shows slightly higher AssA and AssRe due to its two-stage paradigm, DKGTrack maintains a balanced performance in both detection and association tasks, achieving an improvement of 6.63% in HOTA. Compared with some end-to-end methods like TransRMOT [40] and TempRMOT [48], DKGTrack achieves significant improvements, outperforming TransRMOT [40] by 14.09% and TempRMOT [48] by 3.52% in HOTA. These results show the effectiveness of DKGTrack, which decouples language into fine-grained components and leverages them to enhance language understanding and tracking accuracy.

**Refer-KITTI-V2.** We further evaluate DKGTrack on the more challenging Refer-KITTI-V2 dataset, which intro-

Method	Detector	E	HOTA	DetA	AssA	DetRe	DetPr	AssRe	AssPr	LocA
FairMOT [45]	DLA-34	✗	22.53	15.80	32.82	20.60	37.03	36.21	71.94	78.28
ByteTrack [46]	DLA-34	✗	24.59	16.78	36.63	22.60	36.18	41.00	69.63	78.00
iKUN [7]	DeformableDETR	✗	10.32	2.17	49.77	2.36	19.75	58.48	68.64	74.56
TransRMOT [40]	DeformableDETR	✓	31.00	19.40	49.68	36.41	28.97	54.59	82.29	89.82
TempRMOT <sup>‡</sup> [48]	DeformableDETR	✓	33.79	20.89	53.47	31.73	<b>37.37</b>	58.54	82.32	90.20
DKGTrack (Ours)	DeformableDETR	✓	<b>35.26</b>	<b>23.04</b>	<b>54.13</b>	<b>37.81</b>	36.88	<b>60.73</b>	<b>83.85</b>	<b>91.65</b>

Table 2. Quantitative comparison with state-of-the-art MOT methods on the Refer-KITTI-V2 dataset. \* denotes the results after the frame correction. ‡ indicates results reproduced using the official implementation and released weights. The best results are shown in bold.

Base	MPA	SSE	HOTA	DetA	AssA	LocA
✓			50.31	38.49	65.93	90.40
✓	✓		51.07	39.38	<b>66.20</b>	90.43
✓		✓	51.36	40.53	65.58	90.51
✓	✓	✓	<b>52.08</b>	<b>41.10</b>	66.04	<b>90.54</b>

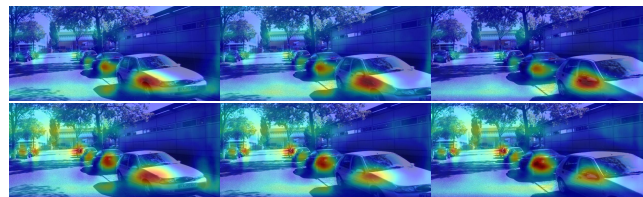
Table 3. Ablation studies on the validation set of the Refer-KITTI dataset to evaluate the contribution of different modules in the DKGTrack. The best results of each metric are marked in bold.

duces more linguistically diverse and semantically rich expressions. As shown in Tab. 2, the proposed DKGTrack consistently outperforms several trackers in both HOTA and DetA. Compared with iKUN [7], DKGTrack achieves significant improvements in both DetA and AssA. These results may be because Refer-KITTI-V2 contains more complex language expressions, making it difficult to simply match the existing tracklets with the entire sentence. Furthermore, DKGTrack achieves an improvement of 4.35% in HOTA and 10.29% in DetA compared with TempRMOT [48]. These results highlight the ability of DKGTrack to effectively comprehend fine-grained knowledge embedded within natural language expressions, enabling it to accurately associate objects with their corresponding language descriptions even in complex scenarios.

#### 4.4. Ablation Study

In this section, we conduct the ablation study to analyze the key components of the proposed DKGTrack, quantifying the individual and combined contributions of each component to the overall tracking performance.

**Analysis of MPA.** We evaluate the effectiveness of the motion perception alignment module, which aligns motion features with object trajectories to enhance temporal consistency during tracking. As shown in Tab. 3, integrating MPA with the baseline improves HOTA from 50.31% to 51.07% and DetA from 38.49% to 39.38%, demonstrating its significant contribution to both detection and tracking accuracy. Notably, the AssA improves to 66.20%, highlighting MPA’s ability to maintain robust object associations over time. However, when MPA is combined with SSE, AssA



Query: “Cars which are parking”

Figure 4. Visualization of the referring regions obtained by TempRMOT (first row) and DKGTrack (second row). While the TempRMOT fails to accurately localize certain cars, DKGTrack effectively focuses on the referred objects, demonstrating its superior localization performance and motion-aware ability.

slightly decreases to 66.04%, indicating that careful balancing of static and motion features is necessary to avoid misalignment between them.

**Analysis of SSE.** We validate the effectiveness of the static semantic enhancement module. As shown in Tab. 3, using the SSE module alone shows a marginal improvement in DetA from 38.49% to 40.53%. This indicates that the SSE module enhances the detection capability of the model by utilizing fine-grained static semantic cues. However, the AssA drops from 65.93% to 65.58%, indicating that while SSE aids in detection, due to lack of motion state information, it is hard to maintain temporal consistency. When combined with MPA, the DKGTrack achieves the best HOTA of 52.08%, demonstrating that SSE and MPA complement each other to improve overall tracking performance by balancing static and motion-based features. Furthermore, Fig. 4 visualizes the heatmap of the Transformer encoder’s final layer output. Compared with TransRMOT [48], which fails to detect distant vehicles in the scene, our method demonstrates superior localization capabilities, accurately identifying and focusing on the targets.

**Analysis of GCE.** As shown in Tab. 4, we analyze the impact of using different language features to initialize object queries in the global context enhancement. Specifically, using entire word-level features shows insufficient for accurate associations, as it fails to distinguish between mixed motion states and localized descriptions, leading to ambiguous target identification. In contrast, decoupling sentence

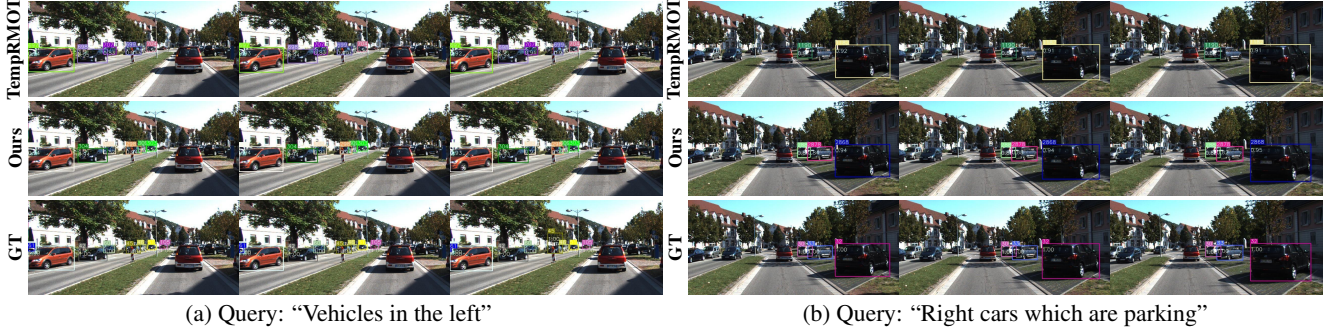


Figure 5. Qualitative comparison between TempRMOT and DKGTracker on the Refer-KITTI dataset. (a) The proposed method exhibits greater discriminative ability for detail descriptions, accurately detecting the white vehicle on the left that is missed by TempRMOT. (b) Compared with TempRMOT, our method is able to correctly track the blue vehicle parking on the right with a high confidence score.

Input	HOTA	DetA	AssA	LocA
$F_W$	50.51	38.49	65.93	90.40
$F_L$	50.87	39.18	<b>66.02</b>	90.43
$F_G$	<b>51.36</b>	<b>40.53</b>	65.58	<b>90.51</b>

Table 4. Ablation studies on using different language features to initialize object queries.

Value	HOTA	DetA	AssA	LocA
0.1	50.33	38.86	65.31	90.30
0.3	50.23	38.29	66.04	<b>90.53</b>
0.5	<b>51.07</b>	<b>39.38</b>	<b>66.20</b>	90.43
0.7	50.78	39.13	65.96	90.25
0.9	50.21	38.14	66.23	90.20

Table 5. Ablation studies on the value of temperature parameter.

embeddings and using localized descriptions alone achieves HOTA of 50.87% and DetA of 39.18%, highlighting the importance of fine-grained language cues. However, this initialization strategy cannot comprehend the full language information, limiting its effectiveness in complex scenarios. Using sentence-level feature achieves the highest HOTA of 51.36%, demonstrating its effectiveness in guiding queries to accurately track referred objects.

**Analysis of Temperature Parameter  $\beta$ .** As shown in Tab. 5, We represent the results of varying the temperature parameter  $\beta$  in the motion perception alignment module.  $\beta$  controls the degree of influence of the motion state descriptions on the object queries. As the value of  $\beta$  increases, the alignment between object tracklets and motion states becomes more accurate, leading to improved tracking performance. When  $\beta$  exceeds 0.5, a declining trend in the evaluation metric is observed. This may be attributed to the excessive fusion between object queries and motion state information, which causes the original features of the target to be damaged, and ultimately leads to the inability to accurately track the relevant targets.

## 4.5. Qualitative Visualization

As shown in Fig. 5, DKGTrack demonstrates exceptional precision in tracking objects described by language expressions, even under challenging conditions such as occlusions and complex motion patterns. In contrast, TempRMOT struggles to understand direction information and fails to detect some referred objects, particularly in scenarios involving occlusions. For instance, TempRMOT fails to track the white vehicle that is partially occluded by the black car on the left, while DKGTrack not only accurately identifies the objects but also maintains robust tracking with higher confidence scores. These results validate DKGTrack’s ability to comprehend language expressions and its robustness in maintaining accurate tracking in complex scenarios.

## 5. Conclusion

In this paper, we propose DKGTrack, a novel referring multi-object tracking method that decouples language expressions into localized descriptions and motion states to enhance visual-language alignment. We propose a static semantic enhancement module to capture the spatial location of referred objects within each frame, effectively improving the detection accuracy. Furthermore, the proposed motion perception alignment module enables the model to comprehend and align object motions, ensuring robust temporal consistency. Experimental results on two challenging benchmarks, Refer-KITTI and Refer-KITTI-V2, demonstrate the effectiveness of DKGTrack, which achieves significant improvements over the previous methods.

## ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grants U21A20514 and 62372388, and in part by the Major Science and Technology Plan Project on the Future Industry Fields of Xiamen City under Grants 3502Z20241027 and 3502Z20241029.

## References

- [1] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *CVPR*, pages 9686–9696, 2023. 2
- [2] Yao Chen, Shuyan Ding, Jianhui Guo, Chen Yang, and Lunbo Li. Cstrack: A comprehensive and concise vision transformer tracker. In *PRCV*, pages 120–132. Springer, 2023. 6
- [3] Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. Sportsmot: A large multi-object tracking dataset in multiple sports scenes. In *ICCV*, pages 9921–9931, 2023. 2
- [4] Qiyuan Dai and Sibe Yang. Curriculum point prompting for weakly-supervised referring image segmentation. In *CVPR*, pages 13711–13722, 2024. 3
- [5] Siyang Dai, Jun Liu, and Ngai-Man Cheung. Referring expression counting. In *CVPR*, pages 16985–16995, 2024. 3
- [6] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *ICCV*, pages 1769–1779, 2021. 3
- [7] Yunhao Du, Cheng Lei, Zhicheng Zhao, and Fei Su. ikun: Speak to trackers without retraining. In *CVPR*, pages 19135–19144, 2024. 2, 6, 7
- [8] Tobias Fischer, Thomas E Huang, Jiangmiao Pang, Linlu Qiu, Haofeng Chen, Trevor Darrell, and Fisher Yu. Qdtrack: Quasi-dense similarity learning for appearance-only multiple object tracking. *IEEE TPAMI*, 45(12):15380–15393, 2023. 2
- [9] Ruopeng Gao and Limin Wang. Memotr: Long-term memory-augmented transformer for multi-object tracking. In *ICCV*, pages 9901–9910, 2023. 3
- [10] Yan Gao, Haojun Xu, Jie Li, Nannan Wang, and Xinbo Gao. Multi-scene generalized trajectory global graph solver with composite nodes for multiple object tracking. In *AAAI*, pages 1842–1850, 2024. 2
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3, 6
- [12] Wenyan He, Yajun Jian, Yang Lu, and Hanzi Wang. Visual-linguistic representation learning with deep cross-modality fusion for referring multi-object tracking. In *ICASSP*, pages 6310–6314. IEEE, 2024. 6
- [13] Yuqing Huang, Xin Li, Zikun Zhou, Yaowei Wang, Zhenyu He, and Ming-Hsuan Yang. Rtracker: Recoverable tracking via pn tree structured memory. In *CVPR*, pages 19038–19047, 2024. 2
- [14] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, pages 6007–6017, 2023. 1
- [15] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 5
- [16] Guorong Li, Hanhua Ye, Yuankai Qi, Shuhui Wang, Laiyun Qing, Qingming Huang, and Ming-Hsuan Yang. Learning hierarchical modular networks for video captioning. *IEEE TPAMI*, 46(2):1049–1064, 2023. 3
- [17] Rui Li, Baopeng Zhang, Jun Liu, Wei Liu, Jian Zhao, and Zhu Teng. Heterogeneous diversity driven active learning for multi-object tracking. In *ICCV*, pages 9932–9941, 2023. 2
- [18] Siyuan Li, Tobias Fischer, Lei Ke, Henghui Ding, Martin Danelljan, and Fisher Yu. Ovtrack: Open-vocabulary multiple object tracking. In *CVPR*, pages 5567–5577, 2023. 2
- [19] Siyuan Li, Lei Ke, Martin Danelljan, Luigi Piccinelli, Mattia Segu, Luc Van Gool, and Fisher Yu. Matching anything by segmenting anything. In *CVPR*, pages 18963–18973, 2024. 2
- [20] Jiacheng Lin, Jiajun Chen, Kunyu Peng, Xuan He, Zhiyong Li, Rainer Stiefelhagen, and Kailun Yang. Echotrack: Auditory referring multi-object tracking for autonomous driving. *IEEE T-ITS*, pages 1–14, 2024. 6
- [21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 6
- [22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 5
- [23] Chang Liu, Xiangtai Li, and Henghui Ding. Referring image editing: Object-level image editing via referring expressions. In *CVPR*, pages 13128–13138, 2024. 3
- [24] Y Liu, M Ott, N Goyal, J Du, M Joshi, D Chen, O Levy, M Lewis, L Zettlemoyer, and V Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 1907. 3, 6
- [25] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, 129: 548–578, 2021. 6
- [26] Weiyi Lv, Yuhang Huang, Ning Zhang, Rucun Song, Mei Han, and Dan Zeng. Diffmot: A real-time diffusion-based multiple object tracker with non-linear prediction. In *CVPR*, pages 19321–19330, 2024. 2
- [27] Yinchao Ma, Yuyang Tang, Wenfei Yang, Tianzhu Zhang, Jinpeng Zhang, and Mengxue Kang. Unifying visual and vision-language tracking via contrastive learning. In *AAAI*, pages 4107–4116, 2024. 3
- [28] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixé, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *CVPR*, pages 8844–8854, 2022. 6
- [29] Sha Meng, Dian Shao, Jiacheng Guo, and Shan Gao. Tracking without label: Unsupervised multiple object tracking via contrastive similarity learning. In *ICCV*, pages 16264–16273, 2023. 2
- [30] Mattia Segu, Bernt Schiele, and Fisher Yu. Darth: Holistic test-time adaptation for multiple object tracking. In *ICCV*, pages 9717–9727, 2023. 2
- [31] Jenny Seidenschwarz, Guillem Brasó, Victor Castro Serrano, Ismail Elezi, and Laura Leal-Taixé. Simple cues lead to a strong multi-object tracker. In *CVPR*, pages 13813–13823, 2023. 2

- [32] Yanyan Shao, Shuting He, Qi Ye, Yuchao Feng, Wenhan Luo, and Jiming Chen. Context-aware integration of language and visual references for natural language tracking. In *CVPR*, pages 19208–19217, 2024. 3
- [33] Yunhang Shen, Chaoyou Fu, Peixian Chen, Mengdan Zhang, Ke Li, Xing Sun, Yunsheng Wu, Shaohui Lin, and Rongrong Ji. Aligning and prompting everything all at once for universal visual perception. In *CVPR*, pages 13193–13203, 2024. 3
- [34] Peize Sun, J Cao, Y Jiang, R Zhang, E Xie, Z Yuan, C Wang, and P Luo. Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020. 6
- [35] Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *CVPR*, pages 20993–21002, 2022. 2
- [36] Pavel Tokmakov, Jie Li, Wolfram Burgard, and Adrien Gaidon. Learning to track with object permanence. In *ICCV*, pages 10860–10869, 2021. 2
- [37] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *ICCV*, pages 3621–3631, 2023. 5
- [38] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, pages 3645–3649, 2017. 6
- [39] Sanghyun Woo, Kwanyong Park, Seoung Wug Oh, In So Kweon, and Joon-Young Lee. Tracking by associating clips. In *ECCV*, pages 129–145, 2022. 2
- [40] Dongming Wu, Wencheng Han, Tiancai Wang, Xingping Dong, Xiangyu Zhang, and Jianbing Shen. Referring multi-object tracking. In *CVPR*, pages 14633–14642, 2023. 1, 3, 6, 7
- [41] Dongming Wu, Tiancai Wang, Yuang Zhang, Xiangyu Zhang, and Jianbing Shen. Onlinerefer: A simple online baseline for referring video object segmentation. In *ICCV*, pages 2761–2770, 2023. 3
- [42] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *CVPR*, pages 4974–4984, 2022. 3
- [43] Jinxia Xie, Bineng Zhong, Zhiyi Mo, Shengping Zhang, Liangtao Shi, Shuxiang Song, and Rongrong Ji. Autoregressive queries for adaptive tracking with spatio-temporal transformers. In *CVPR*, pages 19300–19309, 2024. 2
- [44] Mingzhan Yang, Guangxin Han, Bin Yan, Wenhua Zhang, Jinqing Qi, Huchuan Lu, and Dong Wang. Hybrid-sort: Weak cues matter for online multi-object tracking. In *AAAI*, pages 6504–6512, 2024. 2
- [45] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *IJCV*, 129:3069–3087, 2021. 6, 7
- [46] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *ECCV*, pages 1–21, 2022. 2, 6, 7
- [47] Yujia Zhang, Qianzhong Li, Yi Pan, Xiaoguang Zhao, and Min Tan. Multi-stage image-language cross-generative fusion network for video-based referring expression comprehension. *IEEE TIP*, 33:3256–3270, 2024. 3
- [48] Yani Zhang, Dongming Wu, Wencheng Han, and Xingping Dong. Bootstrapping referring multi-object tracking. *arXiv preprint arXiv:2406.05039*, 2024. 1, 3, 5, 6, 7
- [49] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *ICCV*, pages 2928–2937, 2021. 1
- [50] Li Zhou, Zikun Zhou, Kaige Mao, and Zhenyu He. Joint visual grounding and tracking with natural language specification. In *CVPR*, pages 23151–23160, 2023. 3
- [51] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 6