OmniPT: Unleashing the Potential of Large Vision Language Models for Pedestrian Tracking and Understanding

Teng Fu, Mengyang Zhao, Ke Niu, Kaixin Peng, Bin Li✉

Shanghai Key Laboratory of Intelligent Information Processing
College of Computer Science and Artificial Intelligence, Fudan University
{tfu23, kniu22, kxpeng25}@m.fudan.edu.cn, {myzhao20, libin}@fudan.edu.cn

Abstract

LVLMs have been shown to perform excellently in image-level tasks such as VQA and caption. However, in many instance-level tasks, such as visual grounding and object detection, LVLMs still show performance gaps compared to previous expert models. Meanwhile, although pedestrian tracking is a classical task, there have been a number of new topics in combining object tracking and natural language, such as Referring MOT, Cross-view Referring MOT, and Semantic MOT. These tasks emphasize that models should understand the tracked object at an advanced semantic level, which is exactly where LVLMs excel. In this paper, we propose a new unified **P**edestrian **T**racking framework, namely **OmniPT**, which can track, track based on reference and generate semantic understanding of tracked objects interactively. We address two issues: how to model the tracking task into a task that foundation models can perform, and how to make the model output formatted answers. To this end, we implement a training phase consisting of `RL-Mid Training-SFT-RL`. Based on the pre-trained weights of the LVLM, we first perform a simple RL phase to enable the model to output fixed and supervisable bounding box format. Subsequently, we conduct a mid-training phase using a large number of pedestrian-related datasets. Finally, we perform supervised fine-tuning on several pedestrian tracking datasets, and then carry out another RL phase to improve the model's tracking performance and enhance its ability to follow instructions. We conduct experiments on tracking benchmarks and the experimental results demonstrate that the proposed method can perform better than the previous methods.

Introduction

Pedestrian Tracking (Fu et al. 2023; Cai et al. 2022; Zeng et al. 2022; Zhang, Wang, and Zhang 2023) is a classical task in computer vision, which aims to locate each person in a sequence of images and assign a unique id to each object. In recent years, the task has had a wide range of applications, such as autonomous driving, intelligent surveillance, and sports analytics. However, pedestrian tracking in complex scenarios is still a challenge, and how to stably track an object when it is frequently obscured, blurred, or even disappeared is still a popular and practical topic.

✉ Corresponding author.

Instead, humans can easily keep tracking an object, even if that object disappears after a long period of time. In medicine, this is a joint result of the semantic and situational memory functions in humans (Martin and Chao 2001; Binder and Desai 2011; Kumar 2021). Simply put, we will abstract an object into a semantic description so that we can subsequently perform a subconscious retrieval to recognize the object. This provides new promising ideas for solving the above difficulties that still exist in MOT topics: **Stable semantic information** about the object can be extracted as special appearance features in the traditional sense to assist the tracker in tracking. At the same time, many new multimodal topics have emerged in the MOT field, *i. e.*, tracking based on verbal information (named Referring MOT, RMOT (Wu et al. 2023; Zhang et al. 2024; Fu et al. 2025a)); cross-view object tracking that tracks based on verbal information (named Cross View Referring MOT, CRMOT (Chen, Yu, and Tao 2024)) and semantic multiple object tracking for semantic understanding of tracked objects and videos (named Semantic MOT, SMOT (Li et al. 2025)). These tasks all emphasize that while tracking an object, the model should understand the object at the semantic level.

Meanwhile, LVLMs have demonstrated remarkable performance in image-level understanding, but their capabilities at the instance level and pixel level still have certain gaps compared with previous expert models. In this paper, we propose **OmniPT**, a model that adopts a new "one-for-all" paradigm that not only solves the traditional object tracking problem but also allows referring tracking based on verbal cues or semantic understanding of the tracked object.

We first addressed two issues: how to decompose the tracking task into natural language tasks that LVLMs can perform, and how to make LVLMs output results in a specified format. To this end, based on the pre-trained weights of LVLMs, we implemented a four-stage training strategy: `RL-Mid Training-SFT-RL`. We first applied the GRPO algorithm (Shao et al. 2024) for a very lightweight reinforcement learning training to supervise the model's output format of bounding boxes. Subsequently, we used a large amount of pedestrian-related data and designed some proxy tasks. These tasks lie between the general knowledge learning in the pre-training stage and the specific tasks in the post-training stage, and like many LMs (Wang et al. 2025; OLMo et al. 2024), we refer to this as the Mid Training stage. Then

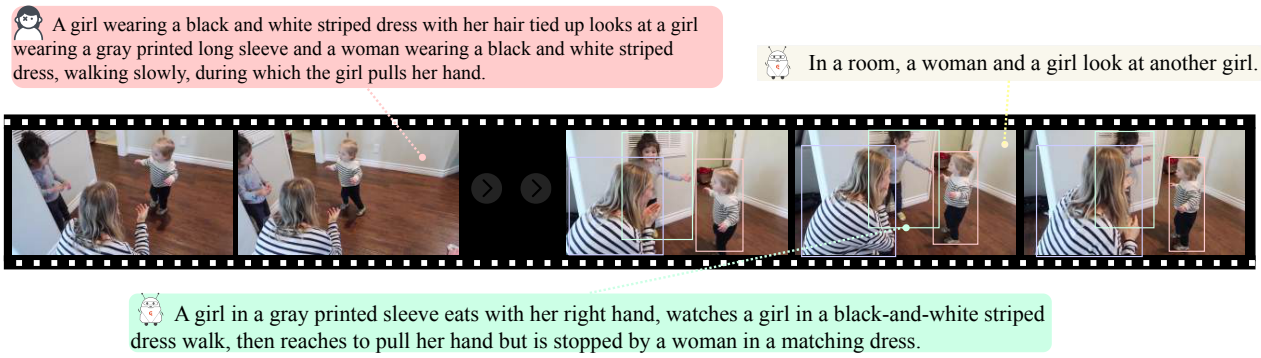

Figure 1: OmniPT can perform traditional object tracking; it can also perform object-specific tracking based on given references and can perform semantic understanding of the tracked objects as well as the whole video.

we performed SFT on multiple pedestrian tracking datasets. At this stage, we decoupled the tracking task into a VQA form to help the model perform the task. Finally, an additional RL stage helped the model further improve its tracking performance and instruction-following ability.

We perform the quantitative analysis of our method on several datasets including DanceTrack (Sun et al. 2022), Refer-KITTI-V2 (Zhang et al. 2024), CRTrack (Chen, Yu, and Tao 2024) and BenSMOT (Li et al. 2025). Most of our selected datasets are pedestrian-specific datasets, which is currently the most popular researched category. However, for RMOT, there is no pedestrian-specific dataset, so we also use these datasets for generalization tests. During test, depending on the task dataset, different instructions can be used to get an output that meets the requirements of the corresponding dataset. Experiments show that our model achieves state-of-the-art results on all datasets with a huge improvement compared to previous best methods. For example, we achieve a HOTA score of 75.04, which is a 3.06 improvement over the previous best result on BenSMOT. Additional ablation experiments also illustrate the effectiveness of the proposed method.

Related Work

Multiple Object Tracking (MOT) aims to locate objects in a video sequence and assign an unique ID to each object across frames. There are several predominant paradigms for MOT, such as Tracking-by-Detection (TBD) (Wojke, Bewley, and Paulus 2017; Cao et al. 2023; Zhang et al. 2021; Fu et al. 2025b) methods and end-to-end methods. TBD methods typically employ detectors to identify objects and associate them by computing positional or appearance-based similarities between detected objects and existing tracklets. SORT (Bewley et al. 2016) is a pioneering approach that integrates deep learning into MOT, leveraging Kalman filtering for motion prediction and the Hungarian matching algorithm for data association. DeepSORT (Wojke, Bewley, and Paulus 2017) further improves upon this by utilizing CNN for appearance feature extraction, combining both positional and appearance similarities for distance matrix calculation. ByteTrack (Zhang et al. 2022) refines the matching process by considering more detection results and employing a two-stage association strategy. Subsequent approaches have focused on enhancing detector performance, refining motion estimation, improving appearance feature extraction, and developing more efficient distance matrix fusion strategies.

The Transformer (Vaswani et al. 2017) is now being widely used in MOT to form a new end-to-end paradigm (Cai et al. 2022; Zeng et al. 2022; Zhang, Wang, and Zhang 2023). Trackformer (Meinhardt et al. 2022) utilizes detection queries and track queries to detect new tracklets and track existing tracklets, respectively. DNMOT (Fu et al. 2023) adopts the "Noising and Denoising" approach to help the model better handle objects in crowded scenarios. Several other paradigmatic approaches have also emerged. DiffMOT (Lv et al. 2024) employs diffusion models to capture object motion between adjacent frames, while graph-based methods (Cetintas, Brasó, and Leal-Taixé 2023; Li, Kong, and Rezatofighi 2022; Dai et al. 2021; Hornakova et al. 2020) represent the relationships between objects as a graph structure and use graph neural networks (GNNs) to associate objects with trajectories.

Language-based Tracking Tasks

In recent years, many new topics in the direction of object tracking have emerged, often involving multi-modality, particularly the intersection of image and natural language modalities. As shown in Figure 2, Referring Multi-Object Tracking (RMOT) (Wu et al. 2023; Zhang et al. 2024) aims to track all objects in a scene that match a given linguistic description. Cross-View RMOT (CRMOT) (Chen, Yu, and Tao 2024), in contrast, performs this task across multiple viewpoints, offering a more comprehensive perspective of the objects and improving the matching between objects and their linguistic references. Semantic Multi-Object Tracking (SMOT) (Li et al. 2025) places greater emphasis on semantic understanding, analyzing both the scenario and the tracked objects. This task consists of four subtasks that extend beyond MOT, including video captioning, instance captioning, and categorizing interactions between different objects.

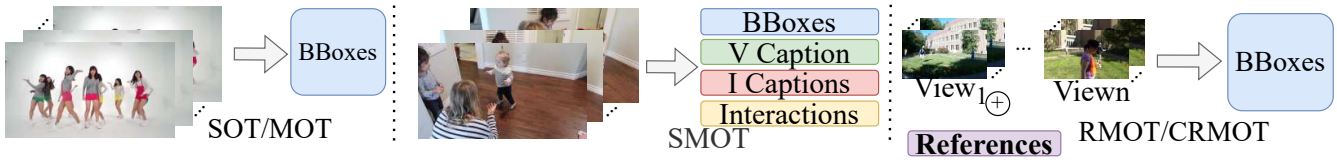

Figure 2: The format of the inputs and outputs for the different tasks. V and I represent Video and Instance.

Foundation model in MOT

Multimodal foundation models have driven MOT research toward tracking a wider variety of objects at a finer level of granularity. SAMTrack (Cheng et al. 2023) leverages SAM (Kirillov et al. 2023) to segment and track all objects in a scene. OVTrack (Li et al. 2023) proposes open-vocabulary MOT, aiming to track all objects in the scene by utilizing CLIP's generalization capabilities for open-world object tracking. MASA (Li et al. 2024b) focuses on fine-grained tracking, exploring instance-level object features using SAM (Kirillov et al. 2023) and detectors such as Grounding DINO (Liu et al. 2024) or YOLOX (Ge et al. 2021). While these methods use LVLMs to track more objects, our approach uses language to guide the model in tracking specific objects.

Methodology

Group Relative Policy Optimization

DeepSeekMath (Shao et al. 2024) proposed GRPO, an effective reinforcement learning algorithm. As shown in Figure 3(b), it optimizes the model towards the direction of meeting ideal preferences by sampling multiple outputs and generating reward scores for these outputs. We use this step to standardize the model's output format of bounding boxes. Specifically, we want the model to use normalized coordinates and strictly follow the `<bbox>x,y,w,h</bbox>` format, where `(x,y)` represents the top-left coordinates and `(w,h)` denotes the size of the bounding box. To this end, we adopt the following reward function:

$$R(bbox_p) = \begin{cases} 2 & bbox_p \in \texttt{set 1}, \\ 0.6 & bbox_p \in \texttt{set 2}, \\ 0.4 & bbox_p \in \texttt{set 3}, \\ 0.0 & otherwise. \end{cases} \quad (1)$$

where `set 1` means the prediction matches the format, `set 2` means that the prediction contains other formats of the bounding box such as `(x,y,x,y)` and `set 3` means that the prediction does not contain `<bbox>...</bbox>` tags. The final reward is computed as follows:

$$R_{s1} = R(bbox_p) \times (\frac{\texttt{IOU}(bbox_p, bbox_{gt})}{2} + 0.5) \quad (2)$$

where `IOU` is IOU score. In the first training stage, we mapped the IOU score to the range of 0.5-1, so that the model would pay more attention to the format at this stage. In the fourth training stage, we canceled this operation. We use the BenSMOT (Li et al. 2025) dataset as the training set in this stage and use the visual grounding as the proxy task.

Mid Training

Traditional object trackers possess three primary capabilities: object detection, position prediction, and ReID. Mid training in this stage aims to enhance the model's sensitivity to language descriptions of tracked objects and improve the model's performance in all three aforementioned aspects. As shown in Figure 3(c), we first train the vision encoder of the baseline model with the image-text aligning task from CLIP (Radford et al. 2021). We adopt the SYNTHPEDES (Zuo et al. 2023) dataset for this training, a large-scale dataset for text-based person re-identification.

Similar to CLIP, we insert special $< CLS >$ tokens into the input sequence and use them to compute similarity scores, forming a similarity matrix supervised using a cross-entropy loss function. Furthermore, we design several pretext tasks to adapt our model to tracking-specific tasks:

Object Detection. For this task, we utilize all available data from the datasets. We randomly sample an image and extract the ground truth bounding boxes as the supervision.

Location Prediction. We sample a segment of the trajectory and provide the object's coordinates in the first frame as input. Using instructional prompts, the model predicts the object's position in the subsequent frame of the trajectory. The ground truth position in the next frame serves as the supervision. Notably, only the object's position in the first frame is given, requiring the model to locate the object in the following frames before making predictions. This implicit training enhances both the model's re-identification and localization capabilities. We also include special samples where the object disappears mid-trajectory, significantly improving the model's robustness. DanceTrack (Sun et al. 2022) is the primary dataset for this training due to its unpredictable object motion patterns.

Person Re-identification. We randomly sample the same person from two different frames and select other people as negative examples to construct a training sample. One frame is used as the anchor sample, and the model is guided to identify the same person from all other samples. To increase diversity, negative samples are drawn both from the same video and from different videos.

Supervised Fine-tuning

We conduct SFT at this stage. Based on the input and output formats of these tasks, we categorize our training samples into four primary types: multi-object tracking, referring tracking, video captioning, and instance captioning.

Multiple Object Tracking. Unlike SOT, MOT requires the detection and initialization of all objects in the first frame, as well as the handling of object disappearance and

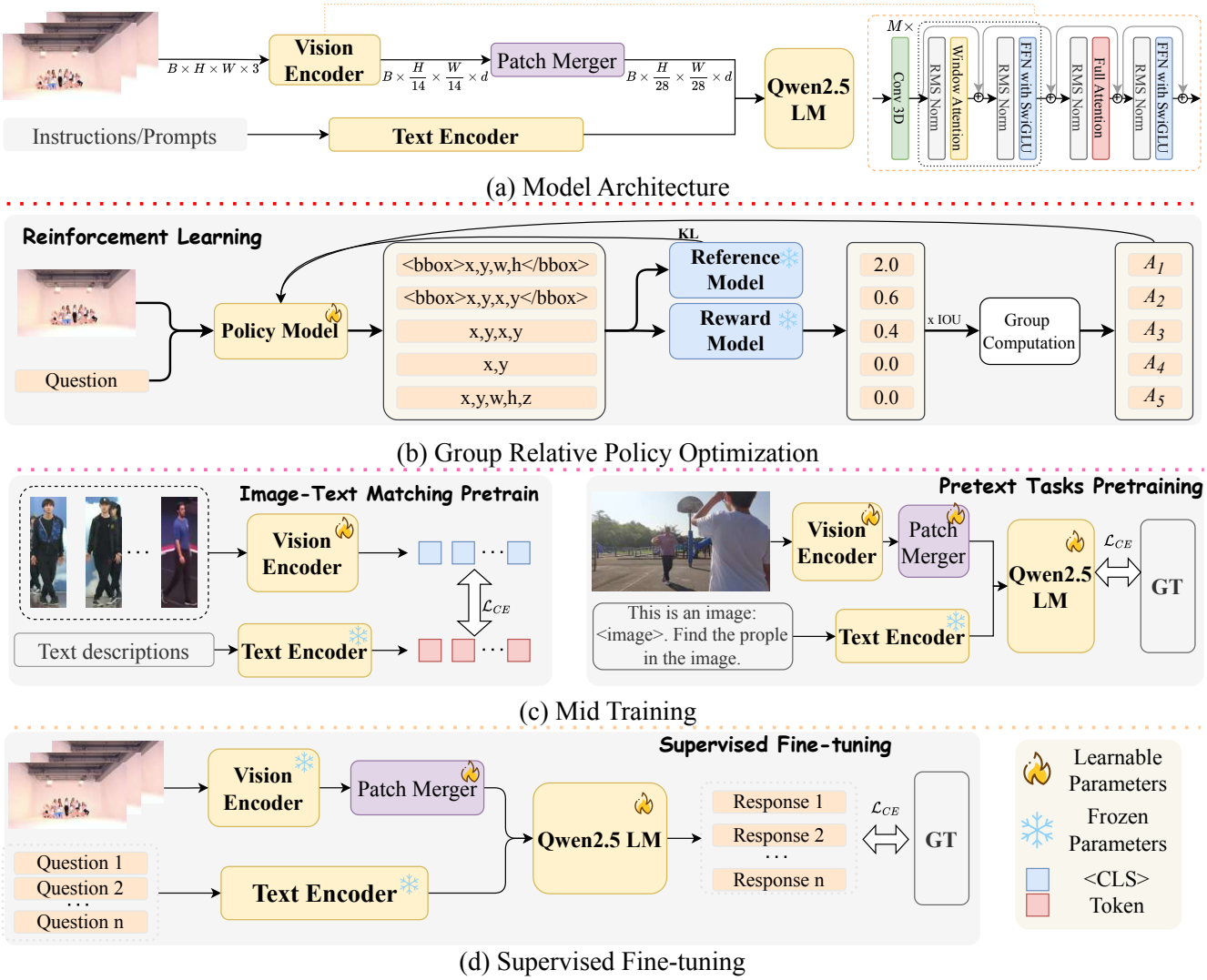

Figure 3: The model architecture of our baseline LVLM and our two-stage fine-tuning strategy for the OmniPT.

the introduction of new objects throughout the sequence. Each training sample consists of a sequence of consecutive video frames and three key queries:

- `Where are the objects located in the first frame?`
- `Track the location of these objects in this sequence.`
- `Are there any objects in the image sequence that were added midway through the sequence?`

The model's response to the first query consists of multiple bounding boxes, presented in random order. The order of these bounding boxes is critical, as it influences the model's supervision of the second query. To ensure efficient training, we specify the true location of each object in the first frame when addressing the second query.

Referring Multiple Object Tracking. RMOT tracks specific objects based on references, differing from MOT by targeting a defined object class and from SOT by tracking all eligible objects rather than a single one. We replace the object position in the first frame in SOT with the linguistic description of the object. CRMOT extends RMOT to multiple views. For each linguistic description, the task involves tracking matching objects across all views simultaneously and assigning consistent IDs to the same object in different views. To address this, we evenly divide the original sequence length among viewpoints and use natural language to explicitly specify the view associated with each sequence.

Video Caption and Instance Caption. These two tasks occur at the end of the object tracking pipeline, offering a semantic understanding of both the tracked object and the video. They can be viewed as captioning tasks, where image sequences and task instructions are provided. In the instance caption task, instead of providing a cropped image of the ob-

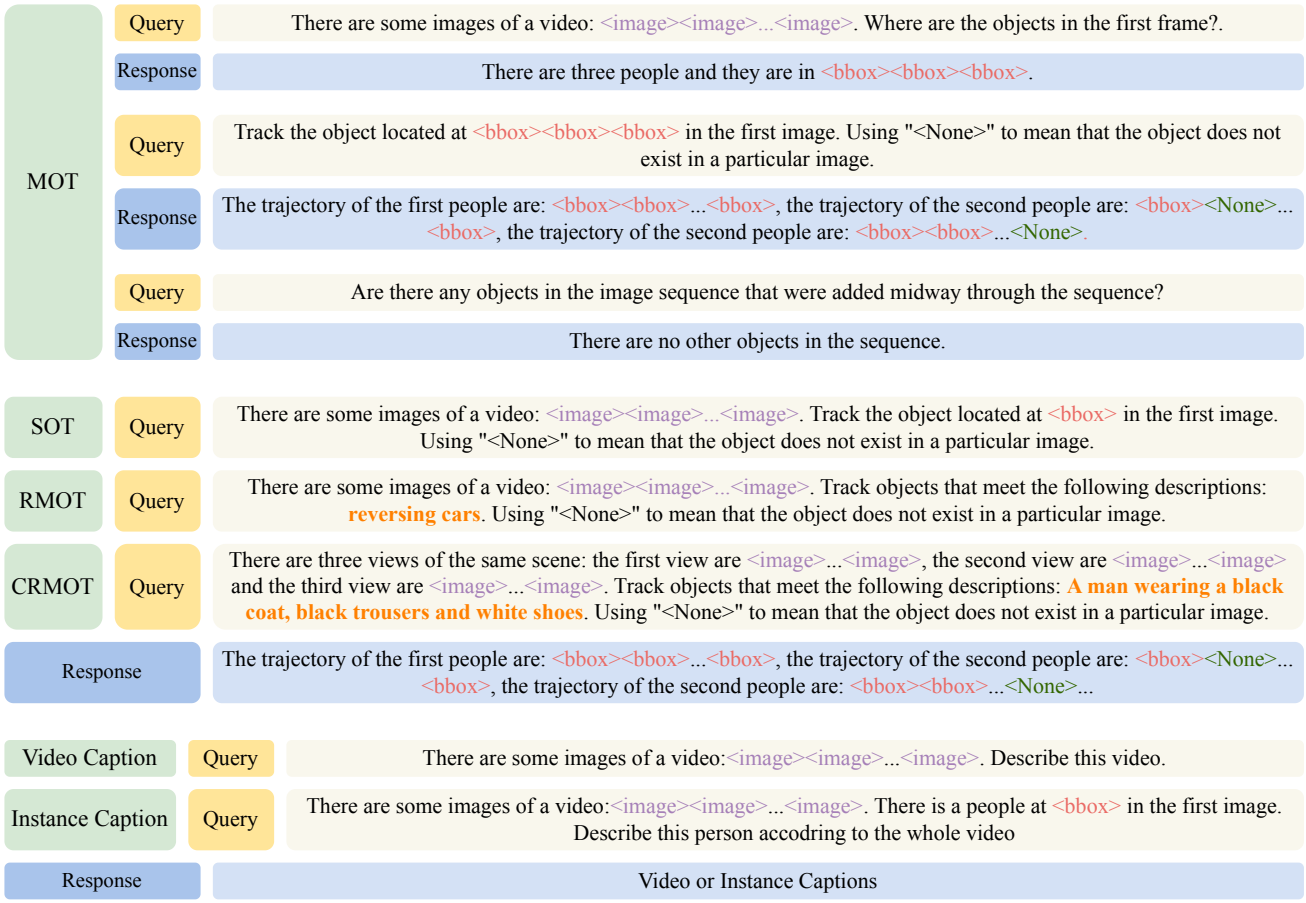

Figure 4: The format of the queries and responses for the different tasks during supervised fine-tuning. RMOT and CRMOT are based on a design for single object tracking.

ject, the entire image is given along with the object's position in the first frame, indicating to the model which specific object in the sequence requires captioning.

Inference

We conduct inference across four tasks: MOT, RMOT, CRMOT, and SMOT. The primary distinction between training and inference lies in the requirement for the model to track objects over extended periods during inference, whereas training involves only sequences within a training sample. To achieve long-term tracking during inference, we employ iterative multi-round dialogues. The tracking result from the last image in the sequence from the previous dialogue round serves as the initial prior information for the new round. For SMOT, which comprises four subtasks, we exclude interaction recognition as it is not related to our research objectives.

Experiments

Datasets and Metrics

MOT. For MOT evaluation, we select DanceTrack (Sun et al. 2022) as our primary benchmark among numerous datasets such as MOT17 (Milan et al. 2016), MOT20 (Dendorfer et al. 2020) and SportsMOT (Cui et al. 2023). The MOT Challenge series, particularly MOT20, often contains densely populated scenes with over 200 people simultaneously, which poses a challenge for our approach due to hardware limitations on the maximum output length in input tokens. While SportsMOT is a large dataset, it lacks comprehensive labeling in certain scenarios (*e.g.*, labeling only players but ignoring spectators on a basketball court), making it less suitable for our purposes. DanceTrack features highly unpredictable object movements and significant appearance similarity between objects, making it an ideal choice for evaluating MOT performance.

SMOT. We evaluate our method using BenSMOT (Li et al. 2025), a dataset comprising 3,292 videos. For each video, we perform tracking on every object and generate captions for both the video as a whole and each individual person within it.

RMOT. We evaluate our model's Referring MOT capability using Refer-KITTI-V2 (Zhang et al. 2024). Compared to its predecessor, Refer-KITTI-V2 expands the data size. Notably, this dataset is not limited to pedestrian tracking but also includes a wide range of objects, such as cars. While

	Video Caption				Instance Caption				Tracking		
Method	BLEU ↑	ROUGE ↑	METEOR ↑	CIDEr ↑	BLEU ↑	ROUGE ↑	METEOR ↑	CIDEr ↑	HOTA↑	IDF1↑	MOTA↑
SORT (Bewley et al. 2016)	0.245	0.224	0.202	0.298	0.233	0.245	0.208	0.056	48.49	53.93	53.58
OC-SORT (Cao et al. 2023)	0.231	0.252	0.215	0.242	0.270	0.205	0.180	0.033	51.00	58.01	55.19
ByteTrack (Zhang et al. 2022)	0.224	0.225	0.212	0.266	0.304	0.242	0.224	0.064	68.84	78.37	73.87
TransTrack (Sun et al. 2020)	0.247	0.248	0.209	0.269	0.283	0.219	0.201	0.074	71.31	78.67	74.08
MOTR (Zeng et al. 2022)	0.187	0.254	0.203	0.244	0.230	0.209	0.182	0.061	66.10	68.97	45.19
MOTRv2 (Zhang, Wang, and Zhang 2023)	0.217	0.258	0.219	0.248	0.238	0.241	0.204	0.059	65.28	70.76	45.52
SMOTer (Li et al. 2025)	0.245	0.261	0.223	0.343	0.306	0.223	0.209	0.087	71.98	80.65	77.71
OmniPT (Ours)	**0.519**	**0.488**	**0.459**	**1.826**	**0.512**	**0.342**	**0.353**	**0.482**	**75.04**	**81.13**	**77.78**

Table 1: Performance comparison between our method and existing methods on the BenSMOT test set. Best results are bolded.

	CRMOT In-domain		CRMOT Cross-domain	
Method	CVRIDF1 ↑	CVRMA ↑	CVRIDF1 ↑	CVRMA ↑
TransRMOT	23.30	8.03	3.66	0.2
TempRMOT	23.43	10.14	3.78	0.39
CRTracker	54.88	35.97	12.52	2.32
OmniPT (Ours)	**62.13**	**42.39**	**46.54**	**33.67**

Table 2: Performance comparison between our method and existing methods on the CRMOT test set.

Method	HOTA ↑	DetA↑	AssA↑	MOTA ↑	IDF1 ↑
FairMOT	39.7	66.7	23.8	82.2	40.8
TransTrack	45.5	75.9	27.5	88.5	45.2
MOTR	54.2	73.5	40.2	79.7	51.5
ByteTrack	47.7	71.0	32.1	89.6	53.9
OC-SORT	55.1	80.3	38.3	**92.0**	54.6
OmniPT (Ours)	**56.4**	**81.7**	**41.0**	90.2	**55.4**

Table 3: Performance comparison between our method and existing methods on the DanceTrack test set.

our primary focus is pedestrian tracking, this dataset allows us to evaluate the generalization performance of our model across diverse object categories.

CRMOT. We evaluate the CRMOT capabilities of our model using CRTrack (Chen, Yu, and Tao 2024), a dataset comprising 13 scenarios. Each scenario is divided into three to four perspectives and includes corresponding linguistic descriptions for the tracked objects.

Metrics. For the MOT, RMOT, and tracking tasks in SMOT, we employ HOTA (Luiten et al. 2021) and CLEAR (Bernardin and Stiefelhagen 2008) as evaluation metrics. For the caption tasks, we adopt ROUGE-L (Lin 2004), BLEU (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), and CIDEr-D (Vedantam, Lawrence Zitnick, and Parikh 2015) to assess performance. For the CRMOT task, we introduce CVRIDF1 and CVRMA (Chen, Yu, and Tao 2024) as our evaluation metrics.

Implementation Details

The proposed method is implemented using PyTorch and trained on 24 NVIDIA A100 GPUs. We apply LoRA-based (Hu et al. 2021) fine-tuning to the parameters, as illustrated in Figure 3. Depending on the number of training samples, we set the training epochs from 3 to 5. The maximum number of image pixels is set to $28 \times 28 \times 646$, approximating a 1920×1080 image downsampled twice. The maximum output length of 2048. We use a `warm up-constant-decay` learning rate strategy with a peak learning rate of 1e-5. During training, we process 16 images per sample (increased to 32 for semantic understanding tasks), randomly sampling images for training and processing video frames sequentially during inference.

Method	HOTA ↑	DetA↑	AssA↑
FairMOT (Zhang et al. 2021)	22.53	15.8	32.82
ByteTrack (Zhang et al. 2022)	24.59	16.78	36.63
iKUN (Du et al. 2024)	10.32	2.17	49.77
TransRMOT (Wu et al. 2023)	31.00	19.40	49.68
TempRMOT (Zhang et al. 2024)	35.04	22.97	53.58
OmniPT (Ours)	**36.15**	**26.68**	**54.62**

Table 4: Performance comparison between our method and existing methods on the Refer-KITTI-v2 test set.

Main Results

Semantic understanding. Table 1 presents the results of our approach for semantic understanding tasks on BenSMOT, including the video caption and instance caption tasks. The results of the baseline methods are sourced from the SMOTer (Li et al. 2025). The experimental results highlight the robustness of our approach, particularly leveraging the strength of LVLM in captioning tasks. Our method achieves improvements of over 50% across all metrics, with the CIDEr metric showing a remarkable fivefold increase.

Multi-object tracking. For the MOT task, Tables 1 and 3 highlight the strong performance of our method. We achieve 75.04 and 56.4 HOTA results on the BenSMOT and DanceTrack datasets, respectively. Notably, on DanceTrack, our method leverages robust location prediction and person re-identification capabilities to effectively distinguish between similar objects, resulting in superior tracking performance.

Referring MOT. Table 4 presents our results on Refer-KITTI-v2. Our approach achieves state-of-the-art performance on the HOTA metric, with significant improvements of 26.68 on DetA and 54.62 on AssA, underscoring the effectiveness of our proposed method. Additionally, we also achieve 75.56 MOTA and 65.30 IDF1.

Cross-view Referring MOT. CRTrack integrates two datasets and is divided into In-domain and Cross-domain setups based on whether the test video originates from the same dataset as the training videos. All methods perform significantly better in the In-domain setting due to the ab-

LVLM	Scale	Instance Caption				Tracking			RMOT
		BLEU ↑	ROUGE ↑	METEOR ↑	CIDEr ↑	HOTA↑	IDF1↑	MOTA↑	HOTA↑
LLaVA-NeXT (Li et al. 2024a)	8B	0.211	0.149	0.172	0.244	49.7	48.6	80.4	26.73
InternVL2.5 (Chen et al. 2024)	4B	0.186	0.173	0.217	0.017	46.5	48.2	77.4	25.02
Qwen2-VL (Wang et al. 2024)	2B	0.370	0.303	0.251	0.256	53.4	51.1	84.3	33.37
Qwen2.5-VL (Bai et al. 2025)	3B	0.379	0.314	0.244	0.263	53.8	51.2	87.4	34.70
Qwen2.5-VL (Bai et al. 2025)	7B	0.446	0.323	0.278	0.316	54.4	52.3	90.1	34.92
Qwen2.5-VL (Bai et al. 2025)	72B	**0.512**	**0.342**	**0.353**	**0.482**	**56.4**	**55.4**	**90.2**	**36.15**

Table 5: Performance comparison between different LVLMs and between different model scale.

Training Stages			Tasks		
MT	SFT	RL	MOT	RMOT	SMOT
	✓		47.38	30.37	0.40
✓	✓		51.89	33.26	0.44
	✓	✓	48.63	35.46	0.41
✓	✓	✓	**56.40**	**36.15**	**0.48**

Table 6: Effect of different stage of training on the final results. MT represent the Mid Training and RL represent the fourth stage of reinforcement learning. We use the HOTA, HOTA, and CIDEr as the metrics for each task.

Length	Video Caption		Instance Caption		Tracking	
	METEOR	CIDEr	METEOR	CIDEr	HOTA	MOTA
2	0.197	0.773	0.149	0.203	52.19	72.36
4	0.258	1.027	0.199	0.271	70.60	76.42
8	0.344	1.370	0.265	0.362	**73.04**	**77.78**
16	0.418	1.774	0.336	0.457	N/A	N/A
32	**0.459**	**1.826**	**0.353**	**0.482**	N/A	N/A

Table 7: Effect of the number of images. N/A represents the inability of the model to handle long inputs.

sence of a domain gap between the training and test sets. Our method achieves a CVRIDF1 of 62.13 and a CVRMA of 42.39 in the In-domain setting, and a CVRIDF1 of 46.54 and a CVRMA of 33.67 in the another setting, demonstrating the effectiveness of our approach in both settings.

Ablation Studies

LVLMs and model scale. We show in Table 5 the effect of different LVLMs and different scales. We choose to use LLaVA-NeXT (Li et al. 2024a), Qwen2 VL (Wang et al. 2024) and InternVL2.5 (Chen et al. 2024) as the comparison methods of our method. We also chose different scales of the QwenVL-2.5 model for the ablation experiments. The Qwen2.5-VL achieved the best results, and we believe that it is the design of the dynamic resolution that leads to a model that can capture the semantic features of the object more clearly, which facilitates the performance of every task.

The different stages of the fine-tuning. We investigate the impact of different stages of fine-tuning in Table 6. Stage 1 is necessary, otherwise we would not be able to monitor the wide variety of outputs in a uniform manner. The results show that gradually increasing pre-training tasks enhance model performance. The first row represents our baseline results **without additional data** to ensure fairness. Mid training stage significantly improves the model's ability to capture pedestrian details, leading to a substantial improvement in tracking performance. RL can further improve the model performance after sufficient training, however, the improvement is very limited when it is not sufficiently trained.

Image length in the training samples. In each training iteration, we provide the model with a set of images for tracking or semantic understanding. Using BenSMOT (Li et al. 2025), we explore the effect of the number of images per iteration on performance. As shown in Table 7, increasing the number of images enhances the model's understanding of objects and scenes. However, for tracking tasks, a smaller number of images leads to more inference rounds, which can weaken the model's memory of the object. Conversely, a larger number of images challenges the model's ability to process multiple images simultaneously. As a result, the overall trend for HOTA and MOTA metrics initially rises and then declines until it cannot be converged.

Limitation

On the one hand, LVLMs struggle to accurately track or even locate all objects in scenes with a large number of objects, such as those in the MOT20 (Dendorfer et al. 2020) dataset. This limitation highlights a critical direction for future advancements in LVLM capabilities. On the other hand, while our work focuses on the "pedestrians" category, we aim to inspire research toward developing unified open-vocabulary object trackers capable of handling diverse object categories.

Conclusion

We propose OmniPT, a novel tracker capable of leveraging object semantic information to perform four tasks simultaneously: MOT, RMOT, SMOT and CRMOT, guided by instructions. Our model is trained through a four-stage training processes on an LVLM baseline. In the first stage, we employed reinforcement learning to standardize the model's output format. Subsequently, we utilized mid-training and supervised fine-tuning to enable the model to learn how to use semantic information for continuous object tracking. Finally, we further enhanced the model's tracking performance and instruction-following ability through reinforcement learning. Extensive experiments on several benchmarks demonstrate our state-of-the-art performance, and ablation studies validate the effectiveness of our approach.

Acknowledgments

This work was supported in part by the National Key R&D Program of China (No.2021ZD0112803), the National Natural Science Foundation of China (No.62176060), and the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning.

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Bernardin, K.; and Stiefelhagen, R. 2008. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008: 1–10.

Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; and Upcroft, B. 2016. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, 3464–3468. IEEE.

Binder, J. R.; and Desai, R. H. 2011. The neurobiology of semantic memory. *Trends in cognitive sciences*, 15(11): 527–536.

Cai, J.; Xu, M.; Li, W.; Xiong, Y.; Xia, W.; Tu, Z.; and Soatto, S. 2022. Memot: Multi-object tracking with memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8090–8100.

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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9686–9696.

Cetintas, O.; Brasó, G.; and Leal-Taixé, L. 2023. Unifying short and long-term tracking with graph hierarchies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22877–22887.

Chen, S.; Yu, E.; and Tao, W. 2024. Cross-View Referring Multi-Object Tracking. *arXiv preprint arXiv:2412.17807*.

Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; et al. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Cheng, Y.; Li, L.; Xu, Y.; Li, X.; Yang, Z.; Wang, W.; and Yang, Y. 2023. Segment and track anything. *arXiv preprint arXiv:2305.06558*.

Cui, Y.; Zeng, C.; Zhao, X.; Yang, Y.; Wu, G.; and Wang, L. 2023. Sportsmot: A large multi-object tracking dataset in multiple sports scenes. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9921–9931.

Dai, P.; Weng, R.; Choi, W.; Zhang, C.; He, Z.; and Ding, W. 2021. Learning a proposal classifier for multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2443–2452.

Dendorfer, P.; Rezatofighi, H.; Milan, A.; Shi, J.; Cremers, D.; Reid, I.; Roth, S.; Schindler, K.; and Leal-Taixé, L. 2020. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*.

Du, Y.; Lei, C.; Zhao, Z.; and Su, F. 2024. ikun: Speak to trackers without retraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19135–19144.

Fu, T.; Chen, Y.; Chen, Z.; Zhao, M.; Li, B.; and Xue, X. 2025a. CrowdTrack: A Benchmark for Difficult Multiple Pedestrian Tracking in Real Scenarios. *arXiv preprint arXiv:2507.02479*.

Fu, T.; Wang, X.; Yu, H.; Niu, K.; Li, B.; and Xue, X. 2023. DeNoising-MOT: Towards Multiple Object Tracking with Severe Occlusions. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2734–2743.

Fu, T.; Yu, H.; Niu, K.; Li, B.; and Xue, X. 2025b. Foundation model driven appearance extraction for robust multiple object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 3031–3039.

Ge, Z.; Liu, S.; Wang, F.; Li, Z.; and Sun, J. 2021. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*.

Hornakova, A.; Henschel, R.; Rosenhahn, B.; and Swoboda, P. 2020. Lifted disjoint paths with application in multiple object tracking. In *International conference on machine learning*, 4364–4375. PMLR.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Kumar, A. A. 2021. Semantic memory: A review of methods, models, and current challenges. *Psychonomic bulletin & review*, 28(1): 40–80.

Li, F.; Zhang, R.; Zhang, H.; Zhang, Y.; Li, B.; Li, W.; Ma, Z.; and Li, C. 2024a. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Li, S.; Fischer, T.; Ke, L.; Ding, H.; Danelljan, M.; and Yu, F. 2023. Ovtrack: Open-vocabulary multiple object tracking. In *Proceedings of the 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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18963–18973.

Li, S.; Kong, Y.; and Rezatofighi, H. 2022. Learning of global objective for network flow in multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8855–8865.

Li, Y.; Li, Q.; Wang, H.; Ma, X.; Yao, J.; Dong, S.; Fan, H.; and Zhang, L. 2025. Beyond MOT: Semantic Multi-Object Tracking. In *European Conference on Computer Vision*, 276–293. Springer.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

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 *European Conference on Computer Vision*, 38–55. Springer.

Luiten, J.; Osep, A.; Dendorfer, P.; Torr, P.; Geiger, A.; Leal-Taixé, L.; and Leibe, B. 2021. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129: 548–578.

Lv, W.; Huang, Y.; Zhang, N.; Lin, R.-S.; Han, M.; and Zeng, D. 2024. DiffMOT: A Real-time Diffusion-based Multiple Object Tracker with Non-linear Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19321–19330.

Martin, A.; and Chao, L. L. 2001. Semantic memory and the brain: structure and processes. *Current opinion in neurobiology*, 11(2): 194–201.

Meinhardt, T.; Kirillov, A.; Leal-Taixe, L.; and Feichtenhofer, C. 2022. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8844–8854.

Milan, A.; Leal-Taixé, L.; Reid, I.; Roth, S.; and Schindler, K. 2016. MOT16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*.

OLMo, T.; Walsh, P.; Soldaini, L.; Groeneveld, D.; Lo, K.; Arora, S.; Bhagia, A.; Gu, Y.; Huang, S.; Jordan, M.; et al. 2024. 2 OLMo 2 Furious. *arXiv preprint arXiv:2501.00656*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Sun, P.; Cao, J.; Jiang, Y.; Yuan, Z.; Bai, S.; Kitani, K.; and Luo, P. 2022. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20993–21002.

Sun, P.; Cao, J.; Jiang, Y.; Zhang, R.; Xie, E.; Yuan, Z.; Wang, C.; and Luo, P. 2020. Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, Z.; Zhou, F.; Li, X.; and Liu, P. 2025. Octothinker: Mid-training incentivizes reinforcement learning scaling. *arXiv preprint arXiv:2506.20512*.

Wojke, N.; Bewley, A.; and Paulus, D. 2017. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, 3645–3649. IEEE.

Wu, D.; Han, W.; Wang, T.; Dong, X.; Zhang, X.; and Shen, J. 2023. Referring multi-object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14633–14642.

Zeng, F.; Dong, B.; Zhang, Y.; Wang, T.; Zhang, X.; and Wei, Y. 2022. Motr: End-to-end multiple-object tracking with transformer. In *European Conference on Computer Vision*, 659–675. Springer.

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 *European Conference on Computer Vision*, 1–21. Springer.

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: 3069–3087.

Zhang, Y.; Wang, T.; and Zhang, X. 2023. Motrv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22056–22065.

Zhang, Y.; Wu, D.; Han, W.; and Dong, X. 2024. Bootstrapping Referring Multi-Object Tracking. *arXiv preprint arXiv:2406.05039*.

Zuo, J.; Yu, C.; Sang, N.; and Gao, C. 2023. Plip: Language-image pre-training for person representation learning. *arXiv preprint arXiv:2305.08386*.