Z-GMOT: Zero-shot Generic Multiple Object Tracking

Kim Hoang Tran[1,2], **Anh Duy Le Dinh**[1], **Tien Phat Nguyen**[1], **Thinh Phan**[3], **Pha Nguyen**[3], **Khoa Luu**[3], **Donald Adjeroh**[4], **Gianfranco Doretto**[4], **Ngan Hoang Le**[3]
[1] FPT Software AI Center, Vietnam
[2] VNUHCM-University of Science, Ho Chi Minh City, Vietnam
[3] Department of Computer Science, University of Arkansas, USA
[4] Department of Computer Science, West Virginia University, USA

Abstract

Despite recent significant progress, Multi-Object Tracking (MOT) faces limitations such as reliance on prior knowledge and predefined categories and struggles with unseen objects. To address these issues, Generic Multiple Object Tracking (GMOT) has emerged as an alternative approach, requiring less prior information. However, current GMOT methods often rely on initial bounding boxes and struggle to handle variations in factors such as viewpoint, lighting, occlusion, and scale, among others. Our contributions commence with the introduction of the *Referring GMOT dataset* a collection of videos, each accompanied by detailed textual descriptions of their attributes. Subsequently, we propose Z − GMOT, a cutting-edge tracking solution capable of tracking objects from *never-seen categories* without the need of initial bounding boxes or predefined categories. Within our Z − GMOT framework, we introduce two novel components: (i) iGLIP, an improved Grounded language-image pretraining, for accurately detecting unseen objects with specific characteristics. (ii) MA − SORT, a novel object association approach that adeptly integrates motion and appearance-based matching strategies to tackle the complex task of tracking objects with high similarity. Our contributions are benchmarked through extensive experiments conducted on the Referring GMOT dataset for GMOT task. Additionally, to assess the generalizability of the proposed Z − GMOT, we conduct ablation studies on the DanceTrack and MOT20 datasets for the MOT task. Our dataset, code, and models are released at: https://fsoft-aic.github.io/Z-GMOT.

1 Introduction

Multiple Object Tracking (MOT) (Bewley et al., 2016; Leal-Taixé et al., 2016; Wojke et al., 2017; Brasó and Leal-Taixé, 2020; Wu et al., 2021; Cao et al., 2023; Maggiolino et al., 2023; Zhang et al., 2022c; Yan et al., 2022; Meinhardt et al., 2022a; Zeng et al., 2022; Cai et al., 2022a) aims to recognize, localize and track dynamic objects in a scene. It has become a cornerstone of dynamic scene analysis and is essential for many important real-world applications such as surveillance, security, autonomous driving, robotics, and biology.

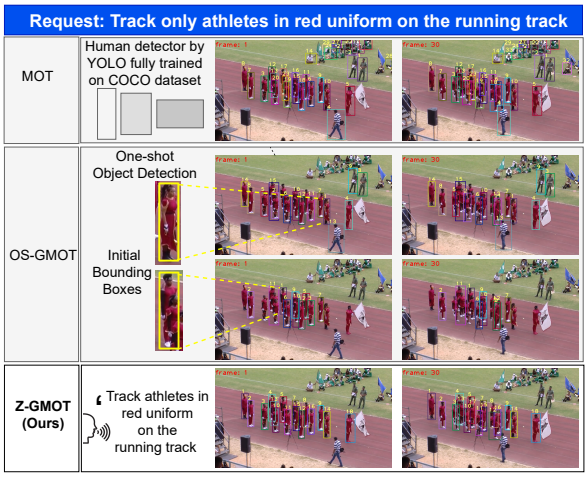


Figure 1: High-level comparison between our Z − GMOT with conventional MOT and one-shot Generic MOT (OS-GMOT) for the task of tracking athletes in red uniforms on a running track. 1^{st} row: MOT, being a fully-supervised method, using YOLOX (trained on COCO) and OC-SORT (trained on DanceTrack) attempts to detect and track all people in the scene with high False Positive (FPs). 2^{nd} row: OS-GMOT is based on an initial bounding box and utilizes an MOT tracker (e.g. OS-SORT in this case). While reducing the number of FPs, OS-GMOT heavily relies on the initial bounding box, leading to variations in results with different bounding boxes and a high number of False Negatives (FNs). 3^{rd} row: our Z − GMOT including: (i) iGLIP effectively detects objects without the need for prior training or initial bounding boxes, and (ii) MA − SORT efficiently associates objects with high visual similarity.

However, current MOT methods suffer from several limitations: they heavily depend on prior knowledge of tracking targets, requiring large labeled datasets; they struggle with tracking objects of unseen or specific categories; they are limited

Findings of the Association for Computational Linguistics: NAACL 2024, pages 3468–3479

in handling objects with indistinguishable appearances. In contrast to MOT, Generic Multiple Object Tracking (GMOT) (Luo and Kim, 2013; Luo et al., 2014) seeks to alleviate these challenges with reduced prior information. GMOT is tailored to track multiple objects of a shared or similar generic type, offering applicability in diverse domains like annotation, video editing, and animal behavior monitoring. Conventional GMOT methods (Luo and Kim, 2013; Luo et al., 2014; Bai et al., 2021) adhere to a one-shot paradigm (Huang et al., 2020) and employ the initial bounding box of a single target object in the first frame to track all objects belonging to the same class. The conventional one-shot GMOT, which is based on one-shot object detection (OS-OD), is known as OS-GMOT. However, this approach heavily relies on the starting bounding box and has limitations in accommodating variations in object characteristics, including pose, illumination, occlusion, scale, texture, etc.

To overcome the aforementioned limitations of both MOT and OS-GMOT, particularly in the context of tracking multiple unseen objects without the requirement for training examples, we introduce a **novel tracking paradigm** called **Zero-shot Generic Multiple Object Tracking** ($\texttt{Z}-\texttt{GMOT}$, which leverages recent advancements in Vision-Language (VL) models. Our $\texttt{Z}-\texttt{GMOT}$ follows the tracking-by-detection paradigm and introduces two significant contributions aimed at enhancing both the object detection stage and object association stage.

In the first stage, which involves object detection, we introduce an enhanced version of GLIP called `iGLIP`. While GLIP has shown promise in detecting objects based on textual description queries, it faces limitations when tasked with detecting multiple objects with subtle distinguishing features. Specifically, our observations and empirical experiments have confirmed that it is sensitive to threshold settings, leading to high False Positives (FPs) at slightly lower thresholds and high False Negatives (FNs) at slightly higher thresholds. For instance, when asked to identify a "red ball" among multiple balls of various colors, GLIP may erroneously detect balls of different colors at a slightly lower threshold and miss the red ball when the threshold is increased only slightly. To address it, our proposed enhancement, `iGLIP`, incorporates two distinct pathways. One pathway is tailored to handle general object categories like "ball", while the other pathway is dedicated to capturing specific object characteristics, such as the color "red". By integrating these dual pathways, `iGLIP` aims to deliver a more accurate and precise object detection process, especially when dealing with multiple generic objects.

In the second stage, which involves object association, we propose $\texttt{MA}-\texttt{SORT}$ (Motion-Appearance SORT), an innovative tracking algorithm that seamlessly fuses visual appearance with motion-based matching. $\texttt{MA}-\texttt{SORT}$ adeptly measures appearance uniformity and dynamically balances the influence of motion and appearance during the association process.

Figure 1 provides a visual comparison between our $\texttt{Z}-\texttt{GMOT}$ with conventional MOT and OS-GMOT approaches with the task of tracking athletes in red uniforms on a running track as an example. In this comparison, MOT is a fully-supervised learning method that employs YOLOX object detection (Ge et al., 2021) trained on COCO dataset (Chen et al., 2015) and OC-SORT object association (Cao et al., 2023) trained on DanceTrack dataset (Sun et al., 2022). Being a fully-supervised method, MOT attempts to detect and track all people in the scene instead of only athletes in red uniforms as requested. As a result, MOT generates a high number of FPs. In contrast, OS-GMOT relies on an initial bounding box to detect all requested objects. It also utilizes the robust OC-SORT tracker (Cao et al., 2023) for object association. While reducing the number of FPs, OS-GMOT heavily relies on the initial bounding box, leading to variations in results with different bounding boxes and a high number of False Negatives (FNs). Different from MOT and OS-GMOT, our $\texttt{Z}-\texttt{GMOT}$ takes the tracking request in the form of a natural language description as its input to effectively detect and track objects without prior training or initial bounding boxes. Our contributions are as follows:

- We introduce a novel tracking paradigm $\texttt{Z}-\texttt{GMOT}$, capable of tracking object categories that have never been seen before, all without the need for any training examples.
- We present *Referring GMOT dataset* consisting of *Refer-GMOT40* and *Refer-Animal* datasets. These datasets are built upon the foundations of the original GMOT-40 dataset (Bai et al., 2021) and the AnimalTrack dataset (Zhang et al., 2022b) with the inclusion of natural language descriptions.
- We propose `iGLIP` to effectively identifies un-

seen objects with specific characteristics.

• We propose MA − SORT, adeptly balancing between object motion and appearance to effectively track objects with highly similar appearances and complex motion patterns.

• We conduct *comprehensive experiments and ablation studies* on our newly introduced Referring GMOT dataset for GMOT task. We extend our experimentation to DanceTrack (Sun et al., 2022), MOT-20 (Dendorfer et al., 2020) datasets for MOT tasks, to illustrate the effectiveness and generalizability of the proposed Z − GMOT framework.

2 Related Works

2.1 Pre-trained Vision-Language (VL) Models

Recent advancements in computer vision tasks have leveraged VL supervision, demonstrating remarkable transferability in enhancing model versatility and open-set recognition. A pioneering work in this domain is CLIP (Radford et al., 2021), which effectively learns visual representations from vast amounts of raw image-text pairs. Since its release, CLIP has garnered significant attention (Yamazaki et al., 2022, 2023; Nguyen et al., 2023; Joo et al., 2023; Yamazaki et al., 2024; Phan et al., 2024; Le et al., 2024; Zhang et al., 2024), and several other VL models, such as ALIGN (Jia et al., 2021), ViLD (Gu et al., 2022), RegionCLIP (Zhong et al., 2022), GLIP (Li et al., 2022b; Zhang et al., 2022a), Grounding DINO (Liu et al., 2023), UniCL (Yang et al., 2022), X-DETR (Cai et al., 2022b), OWL-ViT (Minderer et al., 2022), LSeg (Li et al., 2022a), DenseCLIP (Rao et al., 2022), OpenSeg (Ghiasi et al., 2022), and MaskCLIP (Ding et al., 2022), have followed suit to signify a profound paradigm shift across various vision-related tasks. We can categorize VL pre-training models into three main groups: (i) Image classification: Models in this category, such as CLIP, ALIGN, and UniCL, are primarily focused on matching images with language descriptions through bidirectional supervised contrastive learning or one-to-one mappings. (ii) Object detection: This category encompasses models like ViLD, RegionCLIP, GLIPv2, X-DETR, and OWL-ViT, Grounding DINO, which tackle two sub-tasks: localization and recognition of objects within images. (iii) Image segmentation: The third group deals with pixel-level image classification by adapting pre-trained VL models, including models like LSeg, OpenSeg, and DenseSeg. *In this work, we enhance GLIP and propose* iGLIP *to effectively capture object with specific characteristics.*

2.2 Multiple Object Tracking (MOT)

Recent MOT approaches can be broadly categorized into two types based on whether object detection and association are performed by a single model or separate models, known respectively as joint detection and tracking and tracking-by-detection. In the first category (Chan et al., 2022; Zhou et al., 2020; Pang et al., 2021; Wu et al., 2021; Yan et al., 2022; Meinhardt et al., 2022a; Zeng et al., 2022; Cai et al., 2022a), both objects detection and objects association are simultaneously produced in a single network. In this category, object detection can be modeled within a single network with re-ID feature extraction or motion features. In the second category (Bewley et al., 2016; Leal-Taixé et al., 2016; Wojke et al., 2017; Brasó and Leal-Taixé, 2020; Cao et al., 2023; Zhang et al., 2022c; Nguyen et al., 2022; Aharon et al., 2022; Du et al., 2023; Maggiolino et al., 2023; Cetintas et al., 2023), an object detection algorithm performs detecting objects in a frame, then those objects are associated with previous frame tracklets to assign identities. It is important to note that the state-of-the-art (SOTA) in MOT has been dominated by the later paradigm. Our Z − GMOT approach falls under this paradigm. Particularly, we propose iGLIP for zero-shot objects detector and introduce MA − SORT for objects association generic objects with uniform appearances.

2.3 Generic Multiple Object Tracking (GMOT)

In recent years, MOT has advanced significantly, but it remains tied to supervised learning prior knowledge and predefined categories, complicating the tracking of unfamiliar objects. Different from MOT, GMOT (Luo and Kim, 2013; Luo et al., 2014; Bai et al., 2021) aims to alleviate MOT's limitations by reducing the dependency on prior information. GMOT is designed to track multiple objects of a common or similar generic type, making it suitable for a wide array of applications, ranging from annotation and video editing to monitoring animal behavior. Thus, GMOT often deals with scenarios where objects appear in groups (such as a herd of cows, a school of fish, or a swarm of ants). Consequently, GMOT faces various challenges, including dense object scenarios, small objects, objects with occlusions, among other complexities. Notwithstanding, conventional GMOT methodologies (Luo and Kim, 2013; Luo et al.,

2014; Bai et al., 2021) are predominantly anchored in a one-shot paradigm, i.e. OS-GMOT, leveraging the initial bounding box of a single target object in the first frame to track all objects of the same class. While OS-GMOT shows promise by requiring less prior information, it heavily relies on initial bounding boxes and struggles with viewpoint, lighting, occlusion, and scale variations. Different from MOT (fully-supervised) and OS-GMOT (using initial bounding box), we *introduce a novel zero-shot tracking paradigm known as* Z − GMOT. *Our* Z − GMOT *enables users to track multiple generic objects in videos using natural language descriptors, without the need for prior training data or predefined categories.*

3 Referring GMOT dataset

Table 1: Comparison of **existing datasets** of SOT, MOT, GSOT, GMOT. "#" represents the quantity of the respective items. Cat., Vid. denote Categories and Videos. NLP indicates textual natural language descriptions.

	Datasets	NLP	#Cat.	#Vid.	#Frames	#Tracks	#Boxs
SOT	OTB2013 (Wu et al., 2013)	✗	10	51	29K	51	29K
	VOT2017 (Kristan et al., 2016)	✗	24	60	21K	60	21K
	TrackingNet (Muller et al., 2018)	✗	21	31K	14M	31K	14M
	LaSOT (Fan et al., 2019)	✓	70	1.4K	3.52M	1.4K	3.52M
	TNL2K (Wang et al., 2021)	✓	-	2K	1.24M	2K	1.24M
MOT	MOT17 (Milan et al., 2016)	✗	1	14	11.2K	1.3K	0.3M
	MOT20 (Dendorfer et al., 2020)	✗	1	8	13.41K	3.45K	1.65M
	Omni-MOT (Sun et al., 2020b)	✗	1	-	14M+	250K	110M
	DanceTrack (Sun et al., 2022)	✗	1	100	105K	990	-
	TAO (Dave et al., 2020)	✗	833	2.9K	2.6M	17.2K	333K
	SportMOT (Cui et al., 2023)	✗	1	240	150K	3.4K	1.62M
	Refer-KITTI (Wu et al., 2023)	✓	2	18	6.65K	637	28.72K
GSOT	GOT-10 (Huang et al., 2019)	✗	563	10K	1.5M	10K	1.5M
	Fish (Kay et al., 2022)	✗	1	1.6K	527.2K	8.25k	516K
GMOT	AnimalTrack (Zhang et al., 2022b)	✗	10	58	24.7K	1.92K	429K
	GMOT-40 (Bai et al., 2021)	✗	10	40	9K	2.02K	256K
	Refer-Animal(Ours)	✓	10	58	24.7K	1.92K	429K
	Refer-GMOT40(Ours)	✓	10	40	9K	2.02K	256K

Table 1 presents statistical information for existing tracking datasets including Single Object Tracking (SOT), Generic Single Object Tracking (GSOT), MOT, GMOT. With the recent advancements and the capabilities of Large Language Models (LLMs), there's a growing demand for including textual descriptions in tracking datasets. While natural language have already found their place in SOT and MOT datasets, they have been conspicuously absent from GMOT datasets until now. As a result, our dataset is the pioneering effort to address this demand, integrating textual descriptions into the GMOT domain for the first time.

In this work, we propose to incorporate textual descriptions into two pre-existing GMOT datasets, namely GMOT-40 (Bai et al., 2021) and AnimalTrack (Zhang et al., 2022b), and designate

```
video: stock-3
label:[
  {
   object: "wolf"
   object_synonyms:["wild dog"]
   attributes:["gray fur"]
   other_attributes: ["four legs", "shape
        teeth", "small ear", "strong jaw"]
   tracks: stock-3.text
  }  ]
```



```
video: ball-0
label:[
  {
   object: "ball"
   object_synonyms:["billard ball",
        "sphere", "billard sphere"]
   attributes:["circle", "round", red"]
   other_attributes: ["small", "round"
        "smooth", "numbering", "glossy"]
   tracks: ball_0.text
  }  ]
```

Figure 2: Examples of data annotation structure.

them as the *"Refer-GMOT40"* and *"Refer-Animal"* datasets. *Refer-GMOT40* consists of 40 videos featuring 10 real-world object categories, each containing 4 sequences. *Refer-Animal* contains 26 video sequences depicting 10 prevalent animal categories. Each video undergoes annotation, comprising of an **`object`** name, its corresponding **`attributes`** description, and its corresponding **`tracks`**. It's worth emphasizing that the **`attributes`** description primarily focuses on discernible object characteristics, while **`other_attributes`** aims to offer additional details about the object's traits. Importantly, some of the attributes listed under **`other_attributes`** may not always be visible throughout the entirety of the video. To maintain the standardized format for MOT challenges, as outlined in (Milan et al., 2016; Dendorfer et al., 2020), each video comes with its tracking ground truth, stored in a separate text file within **`tracks`** annotation. This approach ensures consistency with MOT problem conventions. The annotation process follows the JSON format, and Figure 2 offers illustrative examples of the annotation structure. This data is conducted by 4 annotators and made publicly available.

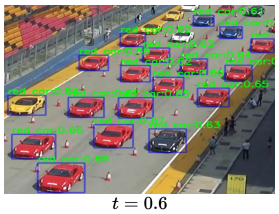
4 Proposed Z − GMOT

Our Z − GMOT framework follows the tracking-by-detection paradigm which includes the object detection stage and object association one. In the initial stage, we analyze the limitations of GLIP detector which is our motivation for proposing iGLIP for detecting effectively generic objects. In the subsequent stage, we introduce MA − SORT to adeptly balance between motion cues and visual appearances to improve the association process.

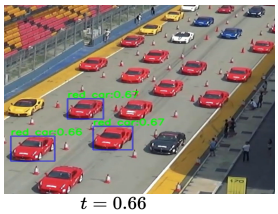
4.1 Proposed iGLIP

We start by analyzing the limitations of GLIP and then proposing iGLIP.

Limitations of GLIP. GLIP encounters difficul-

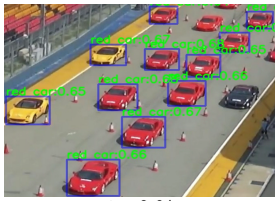


$t = 0.6$

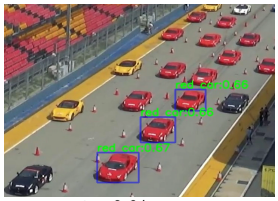


$t = 0.66$

Figure 3: Limitation 1 of GLIP: Sensitive to threshold selection. With slightly different thresholds $t = 0.6$ v.s. $t = 0.66$, GLIP produces different results with high FPs (left) and high FNs (right). Note that GLIP uses prompt *"red car"*) in both results.



$t = 0.64$



$t = 0.64$

Figure 4: Limitation 2 of GLIP: With the same t. ($t = 0.64$) and the same prompt *"red car"*, the results vary when applied to two similar input images.

ties in handling specific object categories (OC^{Spe}) characterized by attributes. As shown in Fig. 3, object detection performance displays sensitivity to threshold selection; even slight threshold variations lead to significant outcome differences. This leads to high TPs at slightly lower threshold and high FNs at slightly higher threshold. Fig. 4 underscores GLIP's drawbacks in effectively capturing objects with specific attributes. Even when using the same threshold selection and prompt, the outcomes exhibit variations on similar images with specific object category OC^{Spe}.

Proposed `iGLIP`. As depicted in Figure 5, our proposed `iGLIP` takes an input image I and two kinds of prompt, namely, a specific prompt (T_s) for OC^{Spe} and a general prompt (T_g) for OC^{Gen}. Both T_s and T_g are derived from the Refering GMOT dataset. Herein, T_g is set as the **`object`** (e.g.,"ball"), while T_s is defined as a combination of **`attributes`** and **`object`** (e.g., "red ball"). Both prompts T_s and T_g go through a text encoder, i.e., BERTModule (Devlin et al., 2018) to obtain contextual word features P_s^0 and P_g^0, respectively. Meanwhile, the image goes through a visual encoder, i.e., Swin (Liu et al., 2021) to obtain proposal features O^0. Then, L deep fusion layers (Li et al., 2022b) are applied into contextual word features P_s^0, P_g^0 and O^0. The i^{th} layer of deep fusion is as follows:

$$O^i_{s-\text{t2i}}, P^i_{s-\text{i2t}} = \text{X-MHA}(O^i_s, P^i_s) \quad (1a)$$

$$O^i_{g-\text{t2i}}, P^i_{g-\text{i2t}} = \text{X-MHA}(O^i_g, P^i_g), \quad (1b)$$

, where specific proposal features are represented by $O_s^{i+1} = \text{DyHeadModule}(O_s^i + O^i_{s-\text{t2i}})$, general proposal features are denoted by $O_g^{i+1} = \text{DyHeadModule}(O_g^i + O^i_{g-\text{t2i}})$, and specific contextual word features $P_s^{i+1} = \text{BERTModule}(P_s^i + P^i_{s-\text{i2t}})$, general contextual word features $P_g^{i+1} = \text{BERTModule}(P_g^i + P^i_{g-\text{i2t}})$. Notably, where L is the number of DyHeadModules in DyHead (Dai et al., 2021) and $O_s^0 = O_g^0 = O^0$. X-MHA denotes a cross-modality multi-head attention module. Finally, the word-region alignment module is utilized to compute the alignment score using dot product between the fused features.

$$S_s^{\text{align}} = O_s P_s^\top, \text{ and } \quad S_g^{\text{align}} = O_g P_g^\top \quad (2)$$

where $O_s = O_s^L \in \mathbb{R}^{N\times d}, O_g = O_g^L \in \mathbb{R}^{N\times d}$ are the visual features from the last visual encoder layer and $P_s = P_s^L \in \mathbb{R}^{M\times d}, P_g = P_g^L \in \mathbb{R}^{M\times d}$ are the word features of OC^{Spe} and OC^{Gen} from the last language encoder layer. The result of this operation are matrices $S_s^{\text{align}} \in \mathbb{R}^{N\times M}, S_g^{\text{align}} \in \mathbb{R}^{N\times M}$. The resulting bounding boxes undergo a filtering process using two parameters: top-κ and threshold $\mathcal{T}$. The top-κ parameter is applied into S_s^{align} to extract a set of queries $\mathcal{B}_q$, which represents template patterns. In order to exclusively detect TPs, we have set $\kappa = 5$. The threshold $\mathcal{T}$ parameter is applied into S_g^{align} to extract a target set $\mathcal{B}_t$. To capture all object proposals, even those potentially including FPs, we set $\mathcal{T} = 0.3$. Query-Guided Matching (QGM) module is then proposed to eliminate FPs in $\mathcal{B}_t$ by using $\mathcal{B}_q$ as template patterns. To perform QGM matching without adding additional cost, we propose to utilize only visual features O^0 extracted from the backbone, without the influence of text embeddings, to ensure the feature is enriched with visual properties. Let O_t^0 and O_q^0 represent the visual features of object proposals in $\mathcal{B}_t$ and $\mathcal{B}_q$, the matching score is defined as the cosine similarity:

$$S_{qt} = \cos(O_q^0 \cdot O_t^{0^T}). \quad (3)$$

The final detection results comprise the query objects and candidate objects with high similarity.

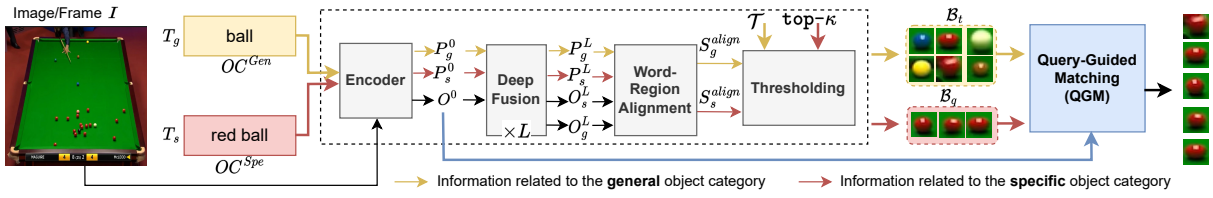


Figure 5: Network architecture of iGLIP, which inputs an image I, a general prompt T_g (e.g. "ball"), and a specific prompt T_s (e.g. "red ball"). iGLIP includes a QGM module to eliminate FPs generated from the general prompt.

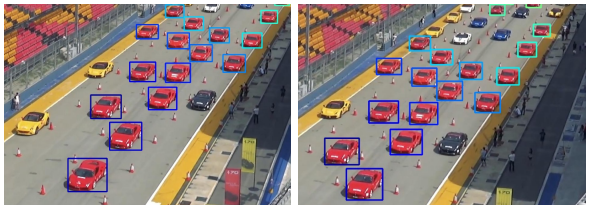


Figure 6: Detection by iGLIP with general prompt T_g as "car" and specific prompt T_s as "red car" across images.

Figure 6 illustrates the red car detection by our proposed iGLIP with general prompt T_g as "car" and specific prompt T_s as "red car" across all images. All results in Figure 6 are generated with the same default settings of $\mathcal{T} = 0.3$ and $\kappa = 5$.

4.2 Proposed MA − SORT

In this section, we introduce our proposed tracking method - *MA-SORT: Balance visual appearance and motion cues:* The standard similarity between N existing track and M detected box embeddings is defined using cosine distance, $C_a \in \mathbb{R}^{M \times N}$. In a typical tracking approach that combines visual appearance and motion cues, the cost matrix C is computed as $C = M_c + \alpha C_a$, where M_c represents the motion cost, measured by the IoU cost matrix. Leveraging DeepOC-SORT (Maggiolino et al., 2023), which computes a virtual trajectory over the occlusion period to rectify the error accumulation of filter parameters during occlusions, the matrix cost becomes:

$$C = IoU + \lambda C_v + \alpha C_a, \tag{4}$$

where C_v represents the consistency between the directions of i) linking two observations on an existing track, and ii) linking tracks' historical observations and new observations. λ and α are hyperparameters to determine the significance of motion and visual appearance, respectively.

To strike a balance between visual appearance and motion cues, we incorporate appearance weight W_a and motion weight W_m into Eq.4. To effectively handle the high similarity between objects of the same generic type in GMOT, we propose the following hypothesis: when the visual appearances of all detections are very similar, the tracker should prioritize motion over appearance. The homogeneity of visual appearances across all detections can be quantified as follows:

$$\mu = \frac{1}{M}\sum_{i=1}^{M} f_i \text{ and } \mu_{det} = \frac{1}{M}\sum_{i=1}^{M} \cos(f_i, \mu). \tag{5}$$

Where M is the number of detections in a frame, f_i is a feature vector of the i-th detection gained from re-ID model (Wojke and Bewley, 2018).

Here, we consider θ as a vector distance threshold to determine the similarity between two vectors; if the angle between them is smaller than θ, the vectors are considered more similar.

$$W_a = \frac{(1 - \mu_{det})}{1 - \cos(\theta)}. \tag{6}$$

We initialize W_m as 1, indicating that both motion and appearance are equally important. As W_a decreases, we propose redistributing the remaining weight to motion, W_m:

$$W_m = 1 + [1 - W_a] = 2 - \frac{(1 - \mu_{det})}{1 - \cos(\theta)}. \tag{7}$$

As a result, the final cost matrix C is:

$$C = W_m(IoU + \lambda C_v) + W_a C_a. \tag{8}$$

5 Experimental Results

5.1 Datasets, Metrics and Experiment Details

We assess our Z − GMOT framework on our Referring GMOT dataset for the GMOT task. To demonstrate the generalizability of Z − GMOT framework, we extend our evaluation to include *DanceTrack* (Sun et al., 2022) and *MOT20* (Dendorfer et al., 2020) for the MOT task. *Refering GMOT dataset*, consisting of *Refer-GMOT40* and *Refer-Animal* dataset, is described in Section 3. *DanceTrack* is

Table 2: Tracking comparison on *Refer-GMOT40* dataset between our iGLIP with SOTA OS-OD (Bai et al., 2021) on various trackers. For each tracker, the best scores are highlighted in **bold**.

Trackers	Detectors	#-Shot	HOTA↑	MOTA↑	IDF1↑
SORT	OS-OD	one-shot	30.05	20.83	33.90
(Bewley et al., 2016)	iGLIP(**Ours**)	zero-shot	**54.21**	**62.90**	**64.34**
DeepSORT	OS-OD	one-shot	27.82	17.96	30.37
(Wojke et al., 2017)	iGLIP(**Ours**)	zero-shot	**50.45**	**58.99**	**57.55**
ByteTrack	OS-OD	one-shot	29.89	20.30	34.70
(Zhang et al., 2022c)	iGLIP(**Ours**)	zero-shot	**53.69**	**61.49**	**66.21**
OC-SORT	OS-OD	one-shot	30.35	20.60	34.37
(Cao et al., 2023)	iGLIP(**Ours**)	zero-shot	**56.51**	**62.76**	**67.40**
Deep-OCSORT	OS-OD	one-shot	30.37	21.10	35.12
(Maggiolino et al., 2023)	iGLIP(**Ours**)	zero-shot	**55.89**	**64.02**	**66.52**
MOTRv2	OS-OD	one-shot	23.75	13.87	25.17
(Zhang et al., 2023)	iGLIP(**Ours**)	zero-shot	**31.32**	**18.54**	**31.28**

Table 3: Tracking comparison on *Refer-GMOT40* dataset between our MA − SORT with other trackers. Our proposed iGLIP is used as the object detection. The best scores are highlighted in **bold**.

Trackers	HOTA↑	MOTA↑	IDF1↑
SORT (Bewley et al., 2016)	54.21	62.90	64.34
DeepSORT (Wojke et al., 2017)	50.45	58.99	57.55
ByteTrack (Zhang et al., 2022c)	53.69	61.49	66.21
OC-SORT (Cao et al., 2023)	56.51	62.76	67.40
Deep-OCSORT (Maggiolino et al., 2023)	55.89	64.02	66.52
MOTRv2 (Zhang et al., 2023)	31.32	18.54	31.28
MA − SORT(**Ours**)	**56.75**	**64.62**	**68.17**

a vast dataset designed for multi-human tracking i.e., group dancing. It includes 40 train, 24 validation, and 35 test videos, totaling 105,855 frames recorded at 20 FPS. *MOT20* is an updated version of MOT17 (Milan et al., 2016) including more crowded scenes, object occlusion, and smaller object size than MOT17.

We employ the following metrics: Higher Order Tracking Accuracy ($HOTA$) (Luiten et al., 2020), Multiple Object Tracking Accuracy ($MOTA$) (Bernardin and Stiefelhagen, 2008), and $IDF1$ (Ristani et al., 2016). $HOTA$ is measured based on Detection Accuracy ($DetA$), Association Accuracy ($AssA$), i.e. $HOTA = \sqrt{DetA \cdot AssA}$, thus, it effectively strikes a balance in assessing both frame-level detection and temporal association performance. All experiments and comparisons have been conducted by an NVIDIA A100-SXM4-80GB GPU.

5.2 Performance Comparison

In Table 2, we benchmark the tracking performance in two scenarios: one involving the use of one-shot object detection (OS-OD) and the other utilizing our proposed zero-shot iGLIP on our newly introduced *Refer-GMOT40* dataset. It is important to note that incorporating OS-OD with these trackers is equivalent to achieving SOTA OS-GMOT (Bai et al., 2021). Table 2 clearly shows that our zero-shot iGLIP, without requiring any prior knowledge or training, achieves significant performance advantages across various metrics when compared to OS-OD, which relies on initial bounding boxes and is run five times. For instance, on OC-SORT tracker, iGLIP shows improvements in HOTA, MOTA, and IDF1 by 26.16, 42.16, and 33.03 points, respectively. On average across all trackers, iGLIP outperforms OS-OD by 21.64, 35.67, and 26.61 points in HOTA, MOTA, and IDF1 metrics.

Table 3 shows the comparison between our proposed MA − SORT with various trackers using the same object detection, i.e., the proposed iGLIP on *Refer-GMOT40* dataset. It is evident that

Table 4: Tracking comparison on *Refer-Animal* between our Z − GMOT and existing *fully-supervised MOT* methods. The best scores are highlighted in **bold**.

Tracker	Detector	Train	HOTA↑	MOTA↑	IDF1↑
SORT	FRCNN(Ren et al., 2015)	✓	42.80	55.60	49.20
DeepSORT	FRCNN(Ren et al., 2015)	✓	32.80	41.40	35.20
ByteTrack	YOLOX(Ge et al., 2021)	✓	40.10	38.50	51.20
TransTrack	YOLOX(Ge et al., 2021)	✓	45.40	48.30	53.40
QDTrack	YOLOX(Ge et al., 2021)	✓	47.00	55.70	56.30
MA − SORT(**Ours**)	YOLOX(Ge et al., 2021)	✓	**57.86**	**68.32**	**63.01**
MA − SORT(**Ours**)	iGLIP (Z − GMOT)(**Ours**)	✗	53.28	57.64	58.43

Table 5: Ablation study of generalizability of Z − GMOT on *DanceTrack* validation set with *MOT task*.

Trackers	Detectors	Train	HOTA↑	MOTA↑	IDF1↑
SORT (Bewley et al., 2016)	YOLOX(Ge et al., 2021)	✓	47.80	88.20	48.30
DeepSORT (Wojke et al., 2017)	YOLOX(Ge et al., 2021)	✓	45.80	87.10	46.80
MOTDT (Chen et al., 2018)	YOLOX(Ge et al., 2021)	✓	39.20	84.30	39.60
ByteTrack (Zhang et al., 2022c)	YOLOX(Ge et al., 2021)	✓	47.10	**88.20**	51.90
OC-SORT (Cao et al., 2023)	YOLOX(Ge et al., 2021)	✓	52.10	87.30	51.60
MA − SORT(**Ours**)	YOLOX(Ge et al., 2021)	✓	**53.44**	87.31	**53.78**
MA − SORT(**Ours**)	iGLIP (Z − GMOT)(**Ours**)	✗	47.57	83.11	46.58

Table 6: Ablation study of effectivess of MA − SORT on *MOT20* testset with *MOT task*. As ByteTrack, OC-SORT (gray) uses different thresholds for test set sequences and offline interpolation procedure, we also report scores by disabling these as ByteTrack†, OC-SORT†. The best scores are highlighted in **bold**.

Trackers	HOTA↑	MOTA↑	IDF1↑
MeMOT (Cai et al., 2022a)	54.1	63.7	66.1
FairMOT (Zhang et al., 2021)	54.6	61.8	67.3
TransTrack (Sun et al., 2020a)	48.9	65.0	59.4
TrackFormer (Meinhardt et al., 2022b)	54.7	68.6	65.7
ReMOT (Fan Yang and Nakamura, 2021)	61.2	77.4	73.1
GSDT (Wang et al., 2020)	53.6	67.1	67.5
CSTrack (Chao Liang and Zou, 2022)	54.0	66.6	68.6
TransMOT (Peng Chu and Liu, 2023)	-	77.4	75.2
ByteTrack(Zhang et al., 2022c)	61.3	77.8	75.2
OC-SORT(Cao et al., 2023)	62.4	75.7	76.3
ByteTrack†(Zhang et al., 2022c)	60.4	74.2	74.5
OC-SORT†(Cao et al., 2023)	60.5	73.1	74.4
MA − SORT(**Ours**)	**61.4**	**77.6**	**75.5**

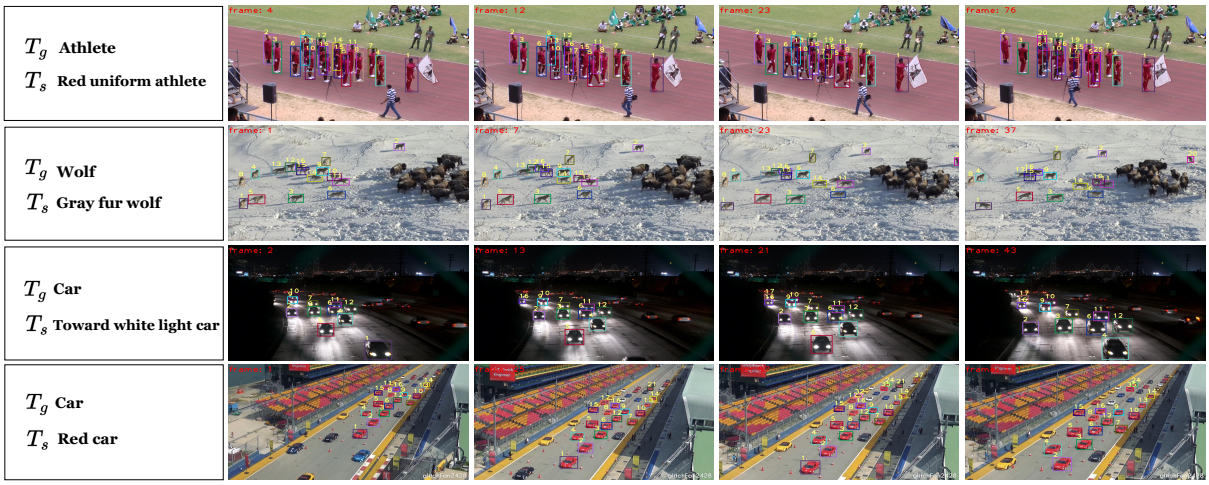


Figure 7: Examples of tracking conducted by our proposed Z − GMOT using input texture descriptions (left). The texture description including both a general prompt T_g, and a specific prompt T_s are integrated into our proposed Z − GMOT framework including iGLIP and MA − SORT.

MA − SORT consistently outperforms other trackers. For example, MA − SORT outperforms DeepSORT and Deep-OCSORT by 6.3, 5.63, 10.62 points and 0.86, 0.6, 1.65 points across all metrics, respectively.

Table 4 presents a comparison between our proposed Z − GMOT and other existing fully-supervised MOT methods on the *Refer-Animal* dataset. In order to ensure a fair comparison, we have also implemented a fully-supervised MA − SORT method with YOLOX object detection. While our MA − SORT with YOLOX object detector achieves the best performance, it is worth noting that Z − GMOT outperforms other fully-supervised MOT methods without the need for any training data.

5.3 Ablation Study

Generalizability of Z − GMOT framework. In addition to the GMOT task, we also evaluate its generalizability on the MOT task, as in Table 5 on DanceTrack dataset. This table presents a comparison of Z − GMOT with the existing fully-supervised MOT methods. To ensure a fair comparison, we have implemented a fully-supervised MA − SORT method with YOLOX object detection. While our MA − SORT with YOLOX achieves the best performance, it is noteworthy that Z − GMOT demonstrates compatibility with SOTA fully-supervised MOT methods, even surpassing SORT, DeepSORT, and MOTDT, all without requiring any training data. In this experiment, both general prompt and specific prompt are set as "dancer".

Effectiveness of proposed MA − SORT. We assess its performance by conducting a comparison on the MOT20 dataset, as outlined in Table 6, focusing on the MOT task. To ensure a fair comparison, we disable certain ad-hoc settings that employ varying thresholds for individual sequences and an offline interpolation procedure. In this experiment, we employed the YOLOX object detector, which demonstrates the effectiveness of MA − SORT.

Effectiveness of proposed iGLIP. We evaluate our iGLIP by comparing it to GLIP (Li et al., 2022b) and OS-OD (Huang et al., 2020) for object detection on the *Refer-GMOT40* dataset, as presented in Table 7(a). iGLIP outperforms other detector methods, achieving the highest scores. It is worth highlighting that despite being an extension of GLIP, iGLIP exhibits significant improvements, with a 0.7% increase in AP_{50}, a 5.0% improvement in AP_{75}, and a 3.9% enhancement in mAP, demonstrating its clear superiority over GLIP.

Table 7: Ablation studies on *Refer-GMOT40*.

(a) Object detection by iGLIP.

Detectors	AP_{50}	AP_{75}	mAP
OS-OD	31.5	13.4	15.8
GLIP	66.2	35.0	36.1
iGLIP	**66.9**	**40.0**	**40.0**

(b) Tracking performance with varied θ.

θ	**HOTA↑**	**MOTA↑**	**IDF1↑**
22.5°	56.57	64.57	67.85
45°	56.58	64.59	67.89
67.5°	56.75	64.62	68.17
80°	56.74	64.62	68.15

Hyper-param θ. Table 7(b) shows ablation study of vector distance threshold θ as defined in Eq.6. The minor variations in tracking performance demonstrate the robustness of our proposed MA − SORT when θ is varied within the range of [22.5°, 80°]. We select $\theta = 67.5$ in the reported results.

5.4 Computational Complexity

To evaluate the computation cost, we report the computational resource of each relevant component and inference time as in Table 8. It is important

Table 8: Properties and computational resources required by our proposed Z-GMOT. Inference time represents an average of 4 videos comprising a total of 1,467 frames.

	MA-SORT + iGLIP				MA-SORT + YOLOX	
Properties						
Settings	**Open-set**				Close-set	
Track Agnotic Objects	✓				✗	
Computational Cost						
	Text Encoder	Vision Encoder	DyHead & RPN	Entire Model	Vision Encoder	Entire Model
Model Size (#Params)	108M	197M	122M	427M	99.1M	124.5M
Inference time (seconds/frame)	0.008	0.019	0.17	0.197	0.064	0.069
FLOPs (G)	45.94	181.32	136.91	364.17	281.9	322
GPU memory (Gb)	1.27	2.84	2.09	6.2	8.6	10.2

Table 9: Comparison of tracking performance and computational complexity between RMOT (Wu et al., 2023) and our MA-SORT with YOLO-X object detection. We report on 2 classes of human and car because RMOT was trained on only those two classes.

Methods	**Tracking Performance**						**Computational Complexity**			
	Human			**Car**			Model size	FLOPs	GPUs Usage	Inference Time
	HOTA	MOTA	IDF1	HOTA	MOTA	IDF1				
RMOT (Wu et al., 2023)	1.075	-0.55	1.19	6.57	2.99	5.41	169M	212G	3 GB	0.118 s/f
MA-SORT + iGLIP	47.02	55.88	52.22	57.8	57.66	71.54	427M	364.17G	6.2 GB	0.197 s/f
MA-SORT + YOLOX	33.08	39.00	41.44	29.88	22.71	34.93	124.5M	322G	10.2 GB	0.069 s/f

to note that the reported inference time represents an average, calculated over 4 videos comprising a total of 1,467 frames. All the implementation and comparison have been conducted on A100 40GB.

In Table 8, we report our computational complexity in two scenarios: (i) open-set setting where the proposed iGLIP is used to detect unseen categories. (ii) close-set setting where YOLOX is used to detect pre-defined class. In both scenarios, we use our proposed MA-SORT as an object association.

To evaluate the effectiveness of our proposed Z-GMOT, we suggest to compare with other state-of-the-art methods in the field, focusing on both computational complexity and performance, as detailed in Table 9. Included in this comparison is RMOT (Wu et al., 2023), a state-of-the-art model in referring-MOT. It is important to note that RMOT is based on fully-supervised learning and operates within a close-set environment, specifically targeting the tracking of persons and cars. The analysis and comparisons presented in Tables 8 and 9 reveal that our Z-GMOT not only holds a comparable computational complexity with state-of-the-art referring tracking methods but also surpasses them with substantial margins.

6 CONCLUSION & DISCUSSION

In this study, we present Z − GMOT, a novel tracking framework capable of tracking diverse objects without relying on labeled data. Z − GMOT adopts a tracking-by-detection paradigm and offers two key contributions: (i) zero-shot iGLIP for effective object detection using natural language descriptions and (ii) MA − SORT for efficient tracking of visually similar objects within a broader context of generic objects. Beyond proposing Z − GMOT, we also introduce a new *Refering GMOT dataset*. We have thoroughly assessed and demonstrated the efficacy and adaptability of Z − GMOT, not only in the GMOT task but also in the MOT task.

Discussion. We utilize GLIP as our preferred VLM for developing iGLIP. However, it is important to recognize the rich diversity of VLMs available in the field, which opens up exciting avenues for deeper exploration. Moreover, in our current study, we have implemented Z − GMOT exclusively using only textual description object and attributes. Nevertheless, our Referring GMOT dataset offers additional information, such as object_synonyms and other_attributes, which hold great potential for further research, particularly in the context of prompt tuning or prompt engineering. Exploring these additional aspects of our Referring GMOT dataset could lead to enhanced object tracking capabilities as well as other fields such as surveillance, robotics, and animal welfare. We expect our work to inspire future research in the unexplored realm of unseen MOT/GMOT paradigms, potentially leading to extensions in other tracking scenarios, e.g., open-vocabulary MOT/GMOT.

References

Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. 2022. Bot-sort: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*.

Hexin Bai, Wensheng Cheng, Peng Chu, Juehuan Liu, Kai Zhang, and Haibin Ling. 2021. Gmot-40: A benchmark for generic multiple object tracking. In *CVPR*.

Keni Bernardin and Rainer Stiefelhagen. 2008. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10.

Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. 2016. Simple online and realtime tracking. In *ICIP*. IEEE.

Guillem Brasó and Laura Leal-Taixé. 2020. Learning a neural solver for multiple object tracking. In *CVPR*.

Jiarui Cai, Mingze Xu, et al. 2022a. Memot: Multi-object tracking with memory. In *CVPR*.

Zhaowei Cai, Gukyeong Kwon, Avinash Ravichandran, Erhan Bas, Zhuowen Tu, Rahul Bhotika, and Stefano Soatto. 2022b. X-detr: A versatile architecture for instance-wise vision-language tasks. *ECCV*.

Jinkun Cao, Jiangmiao Pang, et al. 2023. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *CVPR*.

Orcun Cetintas, Guillem Brasó, and Laura Leal-Taixé. 2023. Unifying short and long-term tracking with graph hierarchies. In *CVPR*.

Sixian Chan, Yangwei Jia, Xiaolong Zhou, Cong Bai, Shengyong Chen, and Xiaoqin Zhang. 2022. Online multiple object tracking using joint detection and embedding network. *Pattern Recognition*, 130.

Yi Lu Xue Zhou Bing Li Xiyong Ye Chao Liang, Zhipeng Zhang and Jianxiao Zou. 2022. Rethinking the competition between detection and reid in multi-object tracking. *IEEE TIP*.

Long Chen, Haizhou Ai, Zijie Zhuang, and Chong Shang. 2018. Real-time multiple people tracking with deeply learned candidate selection and person re-identification. *ICME*, pages 1–6.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. 2023. Sportsmot: A large multi-object tracking dataset in multiple sports scenes. *arXiv preprint arXiv:2304.05170*.

Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. 2021. Dynamic head: Unifying object detection heads with attentions. In *CVPR*.

Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. 2020. Tao: A large-scale benchmark for tracking any object. In *ECCV*. Springer.

Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. 2020. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zheng Ding, Jieke Wang, and Zhuowen Tu. 2022. Open-vocabulary panoptic segmentation with maskclip. *arXiv preprint arXiv:2208.08984*.

Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. 2023. Strongsort: Make deepsort great again. *IEEE TMM*.

Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. 2019. Lasot: A high-quality benchmark for large-scale single object tracking. In *CVPR*.

Sakriani Sakti Yang Wu Fan Yang, Xin Chang and Satoshi Nakamura. 2021. Remots: Self-supervised refining multi-object tracking and segmentation. *Image and Vision Computing*.

Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. 2021. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*.

Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. 2022. Open-vocabulary image segmentation. *ECCV*.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. 2022. Open-vocabulary object detection via vision and language knowledge distillation. *ICLR*.

Lianghua Huang, Xin Zhao, and Kaiqi Huang. 2019. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE TPAMI*, 43(5):1562–1577.

Lianghua Huang, Xin Zhao, and Kaiqi Huang. 2020. Globaltrack: A simple and strong baseline for long-term tracking. In *AAAI*, volume 34.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICLR*. PMLR.

Hyekang Kevin Joo, Khoa Vo, Kashu Yamazaki, and Ngan Le. 2023. Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In *ICIP*, pages 3230–3234. IEEE.

Justin Kay, Peter Kulits, Suzanne Stathatos, Siqi Deng, Erik Young, Sara Beery, Grant Van Horn, and Pietro Perona. 2022. The caltech fish counting dataset: A benchmark for multiple-object tracking and counting. In *ECCV*, pages 290–311. Springer.

Matej Kristan, Jiri Matas, Aleš Leonardis, Tomáš Vojíř, Roman Pflugfelder, Gustavo Fernandez, Georg Nebehay, Fatih Porikli, and Luka Čehovin. 2016. A novel performance evaluation methodology for single-target trackers. *IEEE TPAMI*, 38(11):2137–2155.

Huy Le, Tung Kieu, Anh Nguyen, and Ngan Le. 2024. Waver: Writing-style agnostic text-video retrieval via distilling vision-language models through open-vocabulary knowledge. In *ICASSP*, pages 3025–3029.

Laura Leal-Taixé, Cristian Canton-Ferrer, and Konrad Schindler. 2016. Learning by tracking: Siamese cnn for robust target association. In *CVPRW*.

Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. 2022a. Language-driven semantic segmentation. *ICLR*.

Liunian Harold Li, Pengchuan Zhang, et al. 2022b. Grounded language-image pre-training. In *CVPR*.

Shilong Liu, Zhaoyang Zeng, et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *CVPR*.

Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. 2020. Hota: A higher order metric for evaluating multi-object tracking. *IJCV*, 129(2):548–578.

Wenhan Luo and Tae-Kyun Kim. 2013. Generic object crowd tracking by multi-task learning. In *BMVC*, volume 1.

Wenhan Luo, Tae-Kyun Kim, Bjorn Stenger, Xiaowei Zhao, and Roberto Cipolla. 2014. Bi-label propagation for generic multiple object tracking. In *CVPR*.

Gerard Maggiolino, Adnan Ahmad, Jinkun Cao, and Kris Kitani. 2023. Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification. *arXiv preprint arXiv:2302.11813*.

Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. 2022a. Trackformer: Multi-object tracking with transformers. In *CVPR*.

Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. 2022b. Trackformer: Multi-object tracking with transformers. In *CVPR*.

Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. 2016. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*.

Matthias Minderer, Alexey Gritsenko, et al. 2022. Simple open-vocabulary object detection with vision transformers. *ECCV*.

Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. 2018. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *ECCV*, pages 300–317.

Pha Nguyen, Kha Gia Quach, Chi Nhan Duong, Ngan Le, Xuan-Bac Nguyen, and Khoa Luu. 2022. Multi-camera multiple 3d object tracking on the move for autonomous vehicles. In *CVPR*, pages 2569–2578.

Toan Nguyen, Minh Nhat Vu, An Vuong, Dzung Nguyen, Thieu Vo, Ngan Le, and Anh Nguyen. 2023. Open-vocabulary affordance detection in 3d point clouds. In *IROS*, pages 5692–5698. IEEE.

Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. 2021. Quasi-dense similarity learning for multiple object tracking. In *CVPR*.

Quanzeng You Haibin Ling Peng Chu, Jiang Wang and Zicheng Liu. 2023. Transmot: Spatial-temporal graph transformer for multiple object tracking. In *WACV*.

Thinh Phan, Khoa Vo, Duy Le, Gianfranco Doretto, Donald Adjeroh, and Ngan Le. 2024. Zeetad: Adapting pretrained vision-language model for zero-shot end-to-end temporal action detection. In *WACV*, pages 7046–7055.

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*. PMLR.

Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. 2022. Denseclip: Language-guided dense prediction with context-aware prompting. In *CVPR*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*.

Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. 2016. *Performance Measures and a Data Set for Multi-target, Multi-camera Tracking*, page 17–35.

Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. 2022. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *CVPR*.

Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. 2020a. Transtrack: Multiple-object tracking with transformer. *arXiv preprint arXiv: 2012.15460*.

ShiJie Sun, Naveed Akhtar, XiangYu Song, HuanSheng Song, Ajmal Mian, and Mubarak Shah. 2020b. Simultaneous detection and tracking with motion modelling for multiple object tracking. In *ECCV*, pages 626–643. Springer.

Xiao Wang, Xiujun Shu, Zhipeng Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng Wu. 2021. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In *CVPR*, pages 13763–13773.

Yongxin Wang, Kris Kitani, and Xinshuo Weng. 2020. Joint Object Detection and Multi-Object Tracking with Graph Neural Networks. *arXiv:2006.13164*.

Nicolai Wojke and Alex Bewley. 2018. Deep cosine metric learning for person re-identification. In *WACV*. IEEE.

Nicolai Wojke, Alex Bewley, and Dietrich Paulus. 2017. Simple online and realtime tracking with a deep association metric. In *ICIP*. IEEE.

Dongming Wu, Wencheng Han, Tiancai Wang, Xingping Dong, Xiangyu Zhang, and Jianbing Shen. 2023. Referring multi-object tracking. In *CVPR*, pages 14633–14642.

Jialian Wu, Jiale Cao, Liangchen Song, Yu Wang, Ming Yang, and Junsong Yuan. 2021. Track to detect and segment: An online multi-object tracker. In *CVPR*.

Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. 2013. Online object tracking: A benchmark. In *CVPR*, pages 2411–2418.

Kashu Yamazaki, Taisei Hanyu, Khoa Vo, Thang Pham, Minh Tran, Gianfranco Doretto, Anh Nguyen, and Ngan Le. 2024. Open-fusion: Real-time open-vocabulary 3d mapping and queryable scene representation. *ICRA*.

Kashu Yamazaki, Sang Truong, Khoa Vo, Michael Kidd, Chase Rainwater, Khoa Luu, and Ngan Le. 2022. Vl-cap: Vision-language with contrastive learning for coherent video paragraph captioning. In *ICIP*. IEEE.

Kashu Yamazaki, Khoa Vo, Quang Sang Truong, Bhiksha Raj, and Ngan Le. 2023. Vltint: visual-linguistic transformer-in-transformer for coherent video paragraph captioning. In *AAAI*, volume 37, pages 3081–3090.

Bin Yan, Yi Jiang, Peize Sun, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. 2022. Towards grand unification of object tracking. In *ECCV*.

Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. 2022. Unified contrastive learning in image-text-label space. In *CVPR*.

Fangao Zeng, Bin Dong, et al. 2022. Motr: End-to-end multiple-object tracking with transformer. In *ECCV*. Springer.

Haotian Zhang, Pengchuan Zhang, et al. 2022a. Glipv2: Unifying localization and vision-language understanding. *NIPS*.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *IEEE TPAMI*.

Libo Zhang, Junyuan Gao, Zhen Xiao, and Heng Fan. 2022b. Animaltrack: A benchmark for multi-animal tracking in the wild. *IJCV*.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. 2022c. Bytetrack: Multi-object tracking by associating every detection box. In *ECCV*.

Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. 2021. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *IJCV*, 129:3069–3087.

Yuang Zhang, Tiancai Wang, and Xiangyu Zhang. 2023. Motrv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors. In *CVPR*, pages 22056–22065.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, et al. 2022. Regionclip: Region-based language-image pretraining. In *CVPR*.

Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. 2020. Tracking objects as points. In *ECCV*.