Natural Language Driven Multi-Object Tracking via Joint Spatial, Visual, and Semantic Association

Azhar Hasan
Institute for Software Integrated Systems
Vanderbilt University
azhar.hasan@vanderbilt.edu

Alex Richardson
Institute for Software Integrated Systems
Vanderbilt University
alex.richardson@vanderbilt.edu

Gabor Karsai
Institute for Software Integrated Systems
Vanderbilt University
gabor.karsai@vanderbilt.edu

Abstract

We propose a text-guided multi-object tracking pipeline that couples open-vocabulary object detection with language-driven association to track only objects matching a natural language prompt (e.g., "red van"). The detector generates candidate bounding boxes for any described object, and we introduce new attribute-based filtering and association modules to ensure these tracks remain consistent with the prompt. In particular, a patch-based HSV color voting scheme and scale-aware confidence thresholds filter detections by the described attributes (such as color, position) while retaining small, low-visibility objects. During temporal association, we incorporate spatially masked CLIP embeddings and a text-gated matching cost to enforce semantic alignment between each new detection and the textual description. To evaluate semantic correctness under controlled conditions, we introduce a CARLA-based prompt-tracking benchmark with ground-truth attribute annotations and define prompt compliance metrics. Experiments on KITTI, VisDrone, UA-DETRAC, and the CARLA benchmark demonstrate attribute-selective tracking while maintaining tracking performance comparable to classical class-based trackers.

1. Introduction

Multi-object tracking (MOT) is a central task in computer vision with applications ranging from autonomous driving to aerial surveillance and robotics. Trackers, such as ByteTrack [12], achieve strong performance by coupling motion models with IoU-based association strategies, effectively balancing high and low-confidence detections. However, such approaches remain fundamentally class-driven: they rely on closed sets of predefined categories (e.g., car, pedestrian) and lack the flexibility to detect and track objects described by richer, mission-specific attributes such as "track the red van." The class-based formulation limits their deployment in open-world or mission-critical scenarios where the ability to select and maintain tracks based on natural language descriptions is essential. In contrast, prompt-driven MOT provides the adaptability required for real-world operations. For example, in natural disaster response, drones equipped with such systems could dynamically search for civilians trapped under debris simply by modifying the input prompt—without the need for retraining, which is often time-consuming and power-intensive. By decoupling object detection from fixed categories and enabling natural language-based specifications, this work represents an effort toward generalizing the MOT problem across diverse scenarios, where flexibility and efficiency are paramount. Traditional IoU-based association further exacerbates the challenge in visually crowded or dynamic camera settings. When multiple similar objects appear close together, trackers often confuse their identities, resulting in frequent identifier (ID) switches. While motion cues and bounding box overlap are powerful for temporal association, they provide no mechanism to ensure semantic consistency with an operator's intent. This gap motivates the need for trackers that are both attribute-aware and language-guided. To address these limitations, we propose a text-guided multi-object tracking pipeline that integrates open-vocabulary detection and semantic as-

sociation into a unified framework. Our text-guided MOT approach can flexibly detect and track objects defined by rich natural language, rather than a fixed set of categories. Our pipeline leverages Grounding DINO [5] for free-form, prompt-driven detection and ByteTrack for robust temporal association. Beyond this, we introduce a prompt-consistency module that enforces semantic alignment at association time. Specifically, we use CLIP embeddings [7] and attribute checks (e.g., color, type) to verify that detections remain consistent with the input natural language description throughout a track. This module directly mitigates ID switches by incorporating attribute-level reasoning into data association, an aspect largely absent from existing MOT methods. Our contributions are four-fold:

1. We introduce a text-guided MOT framework that enables selective, prompt-driven tracking in open-world settings.
2. We introduce a lightweight patch-based HSV color voting scheme to verify color attributes in candidate detections, as well as scale-aware adaptive thresholds that lower the confidence requirement for small objects. Together, these steps filter detection outputs to ensure that only objects consistent with the prompt's attributes (e.g., color, type) are considered, improving precision for attribute-specific tracking.
3. We propose a prompt-consistency association module that augments IoU-based matching with semantic similarity checks. Each detection is encoded using a spatially masked CLIP embedding, preserving its context within the full image, to capture both visual and spatial cues. We then incorporate a text-gated matching cost, combining geometry with CLIP-based similarity to the track's reference and the prompt itself.
4. We introduce a controlled CARLA-based prompt-tracking benchmark with ground-truth attribute annotations and propose prompt-compliance metrics to evaluate semantic correctness of tracking with respect to natural language prompts. We validate our pipeline on KITTI, CARLA, VisDrone, and UA-DETRAC, demonstrating competitive tracking accuracy while enabling attribute-selective tracking.

By bridging natural language understanding with multi-object tracking, our approach offers a step toward mission-driven perception systems capable of following targets described in human terms rather than restricted label sets.

2. Related and Motivating Work

Multi-Object Tracking and IoU-Based Association. Classical MOT pipelines follow the tracking-by-detection paradigm, where objects are first localized frame-by-frame and then associated over time. Popular benchmarks such as MOT17 [2] have driven progress in this area, with trackers like DeepSORT [8] incorporating appearance embeddings, and more recently ByteTrack demonstrating that even low-confidence detections can be effectively leveraged via IoU-based matching. However, these methods remain constrained to predefined object classes and rely heavily on geometric cues such as IoU and motion. This reliance makes them vulnerable to identity (ID) switches in crowded scenes or when multiple similar objects are present.

Open-Vocabulary Detection. Recent advances in vision-language pretraining have enabled open-vocabulary detection, allowing detectors to recognize arbitrary concepts described in natural language. Grounding DINO extends the transformer-based DINO detector by tightly integrating text embeddings into its cross-attention layers, enabling precise localization of free-form queries (e.g., "a person on a bike"). Such models mark a shift away from closed label sets and allow flexible, mission-specific detection. However, most open-vocabulary detectors operate at the frame level, without mechanisms for temporal consistency, and thus cannot directly support robust tracking. To address this, we take the detections produced by the Grounding DINO and pass them onto a tracker, enabling the recovery of object tracks over time. This pipeline bridges the gap between open-vocabulary detection and long-term identity preservation.

While Grounding DINO supports free-form natural language prompts, other vision-language models such as Florence-2 [10] adopt more rigid task-specific prompting schemes. Florence-2, for instance, uses tokens such as `<OD>` to indicate object detection tasks, where the input is typically constrained to predefined object categories. Although Florence-2 demonstrates strong performance in zero-shot tasks such as captioning or grounding, its object detection module relies on label-driven formulations and does not support rich, compositional prompts (e.g., "a red van") in the same flexible manner as Grounding DINO. This limits its applicability in scenarios that require nuanced, language-driven tracking.

Language-Guided Tracking. Early attempts to integrate language into visual tracking have focused primarily on video grounding referring to a single object, where a description such as "man in a blue shirt" is localized and followed across frames [11]. While effective for grounding one object, these approaches do not generalize to the multi-object setting, where multiple referents matching the same description may appear simultaneously. The recently proposed Referring Multi-Object Tracking (RMOT) task [9] addresses this gap by intro-

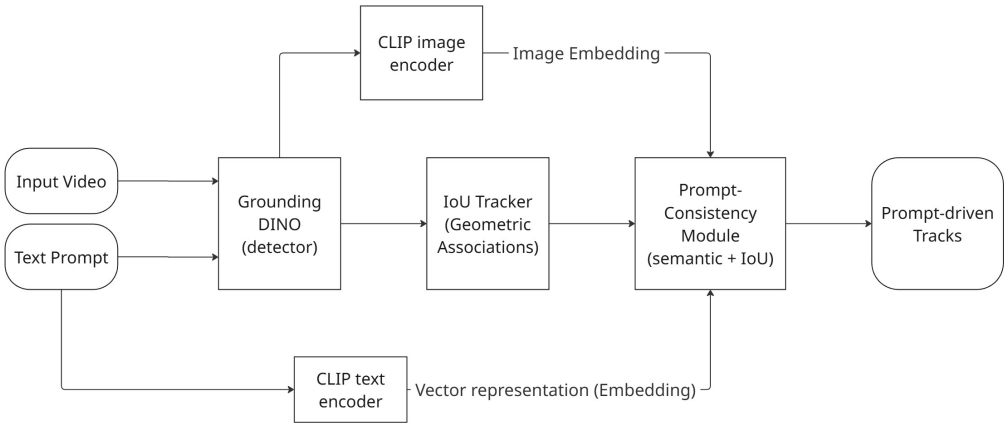

Figure 1. Overview - Given a video and a text prompt, Grounding DINO generates candidate detections. ByteTrack-style association ensures temporal consistency, while a prompt-consistency module combines IoU and CLIP-based semantic similarity to produce final prompt-aligned tracks.

ducing Refer-KITTI, a benchmark with scalable text expressions referring to multiple objects per sequence, and a transformer-based baseline (TransRMOT). This represents the first formalization of multi-object language-guided tracking. However, existing approaches remain tightly coupled to specific architectures and datasets. In contrast, our work takes a modular perspective: we integrate open-vocabulary detection with a classical MOT pipeline, and introduce a prompt-consistency association module that enforces semantic alignment (e.g., color, type) during data association. This makes our system lightweight, extensible, and directly applicable to diverse benchmarks and dynamic camera scenarios, complementing prior end-to-end formulations.

End-to-End Tracking with Attention. An alternative line of work seeks to unify detection and tracking within a single architecture. TrackFormer [6], for instance, formulates multi-object tracking as a set prediction problem, extending the transformer encoder-decoder framework to directly propagate track queries across frames via attention. By eliminating the explicit detection-association split, such models achieve elegant end-to-end formulations. However, this paradigm comes at the cost of modularity: the same model must be retrained whenever the target object categories or task definitions change, making adaptation to new scenarios resource-intensive. In contrast, our approach follows the classical tracking-by-detection paradigm, where detections are obtained independently and then associated temporally and semantically. This separation allows us to flexibly plug in open-vocabulary detectors (e.g., Grounding DINO) and introduce prompt-consistency checks, without retraining the full pipeline. Such modularity makes our system more adaptable to mission-specific requirements (e.g. "civilians in disaster zones") while retaining efficiency and scalability.

3. Approach

3.1. Overview

Our proposed pipeline extends the classical tracking-by-detection paradigm by incorporating language guidance at both the detection and association stages. The overall architecture is illustrated in Fig. 1. Given a video sequence and a natural language prompt (e.g., "red van"), the system first extracts candidate detections using an open-vocabulary detector, then filters and associates them across time with a prompt-consistency module. This enables tracking that is both mission-driven and attribute-aware.

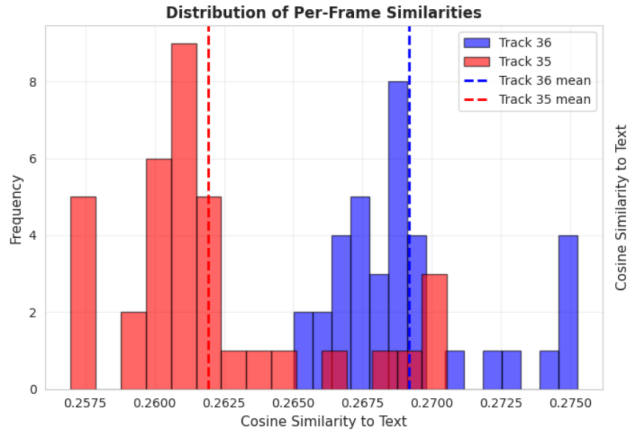

Figure 2. CLIP text–image similarity scores for two vehicle tracks under the prompt "black car on the right." Despite representing visually distinct vehicles, both tracks exhibit highly overlapping similarity values within a narrow range, making reliable color discrimination using CLIP similarity alone difficult.

3.2. Detection

Given a frame I_t and a natural language query q, we use Grounding DINO to obtain candidate detections:

$$\mathcal{D}_t = \{(b_i, s_i, f_i)\}_{i=1}^{N}, \tag{1}$$

where b_i are bounding boxes, s_i are confidence scores, and f_i are text-aligned visual features. Grounding DINO integrates query embeddings into its cross-attention layers, enabling detection of arbitrary concepts. The language prompts are encoded using a pre-trained BERT encoder [3] (bert-base-uncased), which produces token-level embeddings that are fused with visual features through attention. However, to further refine these proposals, we apply two techniques:

- **Patch-Based HSV Color Voting:** If the prompt includes a specific color attribute, we perform an explicit color consistency check on each detection. We subdivide the bounding box region into patches and compute the dominant color in each patch in HSV color space. Each detection then accumulates "votes" from its patches for matching the target color. We retain the detection only if a sufficient fraction of its patches exhibit the specified color, indicating that the object's overall color aligns with the prompt.
  While CLIP embeddings encode high-level semantic appearance, they are not optimized for fine-grained color discrimination, particularly under illumination changes and specular highlights common in vehicle imagery. In our preliminary analysis, we observed that CLIP-based text–image similarities for color attributes tend to collapse into a narrow numerical range of 0.2–0.3 cosine similarity shown in Fig. 2, with visually distinct colors such as white and black vehicles exhibiting highly overlapping similarity scores. This limited dynamic range makes it difficult to reliably threshold or rank detections by color using CLIP alone. In contrast, a simple HSV-based patch voting scheme provides a lightweight and interpretable mechanism for enforcing color consistency at the pixel level, complementing CLIP's strength in capturing higher-level semantic and contextual cues.
- **Scale-Aware Confidence Thresholding:** Small objects often receive lower confidence scores from the detector, making them prone to being discarded by a fixed threshold. To address this, we adopt an adaptive confidence threshold that accounts for the object's scale. In practice, we lower the detection confidence cutoff τ for bounding boxes below a certain pixel area (relative to the frame size). This scale-aware threshold ensures that valid small objects described by the prompt are not overlooked due to conservative confidence settings. By dynamically adjusting τ based on object size, we improve recall for tiny targets while still pruning out truly low-confidence spurious detections.

3.3. Temporal Association

We extend ByteTrack for data association. Given detections $\mathcal{D}_t$ and existing tracks $\mathcal{T}_{t-1}$, associations are formed by maximizing

$$\max_{i,j} \ \mathrm{IoU}(b_i, b_j) \quad \text{s.t.} \ \ s_i > \tau, \tag{2}$$

where b_i denotes a detection box with confidence s_i and b_j is the predicted track box. Detections with $s_i < \tau$ are recovered in a second pass, allowing even low-confidence detections to be associated. This provides robust geometric matching but remains prone to ID switches when visually similar objects overlap.

3.4. Prompt-Consistency Module

Our approach introduces two main components: a spatially masked CLIP-based detection embedding and a text-gated matching cost.

Spatially Masked CLIP Embeddings: We embed each candidate detection using CLIP's vision encoder with the full image context by masking out irrelevant regions, everything outside the detection's bounding box. For each detection b_i, we extract a normalized embedding $e_i \in \mathbb{R}^d$. Each active track t_j maintains a reference embedding e_j^{ref}, updated via an exponential moving average.

Text-Grounded Matching Cost: We introduce an

additional term derived from the textual description to steer the association. First, we compute a CLIP text embedding e_t for the prompt q using CLIP's text encoder. During association, for each candidate match between detection i and track j, we combine three factors: IoU overlap, visual embedding similarity, and text embedding similarity. The association score is defined as:

$$S(i,j) = (1-\lambda_v)\cdot \mathrm{IoU}(b_i, \hat{b}_j) + \lambda_v \cdot \cos(e_i, e_j^{\mathrm{ref}}) + \lambda_t \cdot \mathcal{T}(e_i, e_t), \tag{3}$$

where:

- $\mathrm{IoU}(b_i, \hat{b}_j)$ denotes the spatial overlap between detection i and the predicted position of track j,
- $\cos(e_i, e_j^{\mathrm{ref}})$ is the cosine similarity between the detection's appearance embedding e_i and the reference embedding e_j^{ref} of track j,
- $\mathcal{T}(e_i, e_t) \in [0, 1]$ is the text-grounding cost, computed as the dissimilarity between detection embedding e_i and the prompt embedding e_t,
- λ_v and λ_t are weighting coefficients that balance appearance and text terms, respectively.

This formulation implements a three-way cost fusion. The parameter $\lambda_v \in [0, 1]$ balances spatial and appearance terms. Association is solved using the Hungarian algorithm on $-S(i, j)$. Unlike classical appearance-based trackers that rely solely on image–image similarity, our formulation introduces explicit text grounding as a first-class association term. This allows association decisions to be conditioned not only on visual continuity, but also on semantic alignment with the operator's intent. Importantly, the text-grounding term acts as a stabilizer when IoU and appearance cues are ambiguous, such as in crowded scenes with visually similar objects.

4. Experimental Setup

4.1. Datasets

We evaluate our text-guided multi-object tracking pipeline on three publicly available benchmarks: VisDrone, UA-DETRAC [4], Kitti and Carla. These datasets present varied real-world tracking scenarios, including aerial surveillance, urban traffic monitoring, and multi-class pedestrian-vehicle scenes.

- **VisDrone** contains high-density aerial footage captured from drones, featuring significant occlusions, scale variation, and frequent object interactions. This dataset is used for both training and evaluation to assess our pipeline's performance in visually complex, dynamic scenes.
- **UA-DETRAC** and **KITTI** are used strictly for zero-shot evaluation, without any additional training or fine-tuning. These benchmarks test the generalization capabilities of our model using free-form textual prompts, relying entirely on the open-vocabulary nature of Grounding DINO and the semantic consistency provided by our Prompt-Consistency module.
- **Carla** In addition to real-world benchmarks, we introduce and evaluate on a controlled CARLA-based dataset generated in simulation. Each sequence contains prompt-valid targets and distractors with ground-truth bounding boxes, identities, and attribute annotations (e.g., object type and color). This dataset enables evaluation of prompt-compliance metrics, which cannot be reliably computed on public MOT datasets due to the absence of attribute-level ground truth.

4.2. Implementation Details

Our text-guided MOT pipeline is implemented in PyTorch, integrating the Grounding DINO base model with a Swin-B (Swin Transformer - Base) backbone, configured with an image size of 384×384 and 22k pretrained weights [5]. We follow the classical tracking-by-detection paradigm, combining open-vocabulary object detection with temporal association via ByteTrack, and introducing a Prompt-Consistency Module based on CLIP for semantic alignment. We use a pretrained BERT-base model for prompt encoding, consistent with Grounding DINO's language branch, a component responsible for converting natural language descriptions into representations that can be matched with visual features.

We fine-tune the model on the VisDrone dataset for 20 epochs using the AdamW optimizer. The base learning rate is set to $2.0e{-}4$, and a multi-step decay schedule reduces it by a gamma factor of $(0.1 * baselearningrate)$ at epochs 10 and 15. The visual backbone and text encoder layers (specifically backbone.0 and BERT) are frozen, while projection layers receive a lower learning rate of $1.0e-5$.

To fine-tune thresholds, we conducted hyperparameter optimization on VisDrone using Optuna [1]. The best configuration discovered was: box threshold 0.4, text threshold 0.8, track threshold 0.45, match threshold 0.85, and track buffer of 80. This setting produced the highest MOTA on the validation set.

4.3. Metrics

Following the MOTChallenge protocol [2], we report standard tracking metrics including MOTA, IDF1, and ID switches (IDsw). While MOTA emphasizes detection quality, IDF1 and IDsw measure identity preservation, which is the primary focus of multi-object tracking.

However, these metrics do not capture whether pre-

Table 1. Quantitative comparison with ByteTrack on VisDrone, KITTI (zero-shot), and UA-DETRAC (zero-shot). While overall MOTA remains comparable, our method enables prompt-driven selectivity and semantic consistency during tracking, which is not captured by class-based baselines.

Dataset	Method	MOTA↑	MOTP↓	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDsw↓
VisDrone	Ours (CLIP Tracker)	40.23	0.2111	57.43	221	226	1892	5764	1514
	ByteTrack (baseline)	40.42	0.2094	57.73	211	237	1819	5840	1450
KITTI (zero-shot)	Ours (CLIP Tracker)	46.35	0.231	67.37	51	37	4512	6085	152
	ByteTrack (baseline)	48.78	0.229	67.61	45	42	3914	6504	118
UA-DETRAC (zero-shot)	Ours (CLIP Tracker)	53.74	0.145	75.68	1459	175	132953	171373	125
	ByteTrack (baseline)	53.75	0.145	75.70	1459	175	171435	171435	121

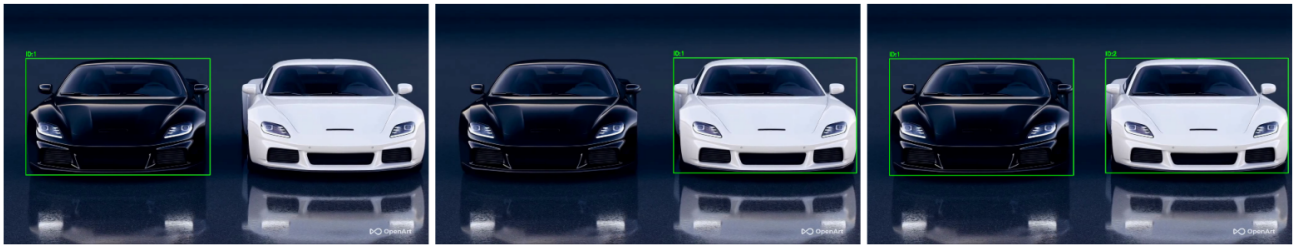

Figure 3. Prompt-driven tracking behavior under different textual queries. The same video sequence is processed with prompts “black car”, “white car”, and “car”, respectively. The system selectively instantiates and maintains tracks consistent with the prompt, demonstrating that semantic constraints influence both detection filtering and temporal association.

dicted tracks satisfy the semantic constraints specified by a natural language prompt. Since our objective is prompt-conditioned tracking, we additionally report a set of prompt-compliance metrics designed to measure semantic correctness.

Let P denote the set of predicted tracks and G the set of ground-truth objects. For each frame t, predicted boxes are matched to ground-truth objects using Hungarian assignment with an IoU threshold θ (0.5 in our experiments). Given a prompt q, we define a predicate

$$\phi_q(g) \in \{0, 1\}$$

which evaluates whether a ground-truth object satisfies the prompt attributes.

Semantic Precision (SP)

$$SP = \frac{\text{Predictions matched to prompt-valid objects}}{\text{Total predictions}}$$

Semantic Recall (SR)

$$SR = \frac{\text{Prompt-valid GT boxes matched by predictions}}{\text{Total prompt-valid GT boxes}}$$

Prompt Coverage Ratio (PCR)

$$PCR = \frac{\text{Total matched prompt-valid GT across frames}}{\text{Total prompt-valid GT across frames}}$$

Distractor Confusion Rate (DCR)

$$DCR = \frac{\text{Predictions matched to prompt-invalid objects}}{\text{Total predictions}}$$

Semantic ID Switches (SID) A semantic ID switch is counted when a predicted track transitions between a prompt-valid and prompt-invalid object in consecutive frames:

$$\phi_q(g_t) \neq \phi_q(g_{t-1})$$

5. Results

5.1. Quantitative Results

Table 1 summarizes quantitative results across three benchmarks, comparing our prompt-consistent tracker against the standard ByteTrack baseline. Across all datasets, our method achieves comparable tracking accuracy while enabling language-driven selectivity. A key advantage of vision-language models in tracking is the ability to perform multi-object tracking under a shared semantic prompt while simultaneously enabling selective tracking through more specific prompts. For example, a generic prompt such as ”car” allows the system to track all vehicles of that category, functioning similarly to class-based MOT. In contrast, a more specific prompt such as ”black car” or ”red sedan” restricts tracking

to semantically consistent targets without modifying the detector or retraining the tracker. This capability allows the same tracking pipeline to support both conventional multi-object tracking and targeted tracking of specific object subsets, providing flexibility not available in fixed-class tracking frameworks. Figure 3 demonstrates that a single tracking system can operate in both category-level and attribute-level modes depending on the prompt.

VisDrone Results. Using the prompt *"car. pedestrian."*, our method achieves a MOTA of 40.23% and IDF1 of 57.43% on the VisDrone validation set. Compared to the ByteTrack baseline (MOTA of 40.42%, IDF1 of 57.73%), our CLIP-enhanced tracker shows comparable performance.

KITTI (Zero-Shot). On the KITTI benchmark in a zero-shot setting, our approach achieves a MOTA of 46.35% and IDF1 of 67.37%, compared to 48.78% MOTA and 67.61% IDF1 for ByteTrack. Although ByteTrack achieves slightly higher aggregate accuracy, our method maintains comparable identity consistency while operating under open-vocabulary conditions. This result highlights the ability of our tracker to generalize across datasets without retraining, leveraging language guidance to preserve semantic consistency during association.

UA-DETRAC (Zero-Shot). On UA-DETRAC, both methods achieve nearly identical MOTA (53.74% for ours and 53.75% for ByteTrack) and IDF1 (75.68% vs. 75.70%).

Prompts and Frame Rate. All experiments are conducted with a frame rate of 24 FPS, except KITTI which works on 10 FPS. We note that prompt granularity (e.g., separating *"car"* and *"pedestrian"*) enables selective tracking across categories without retraining.

6. Experimental Evaluation

6.1. Controlled CARLA Prompt-Tracking Benchmark

To evaluate prompt compliance under controlled conditions with unambiguous ground truth, we introduce a CARLA-based benchmark. Each clip contains a single uniquely identifiable target object specified by an attribute prompt, along with multiple distractors. We design scenarios along multiple axes of difficulty, including distractor similarity, illumination and weather variation, camera motion and viewpoint, scene density, and long-sequence tracking conditions.

CARLA provides ground-truth track identities, per-frame bounding boxes, and semantic attributes (object type and color), enabling direct measurement of prompt compliance without manual annotation.

The prompt-compliance metrics defined in Section 4.3 (SP, SR, PCR, DCR, and SID) are computed only on this handcrafted CARLA benchmark, since publicly available MOT datasets do not provide reliable attribute-level ground truth required to evaluate semantic correctness with respect to a natural language prompt.

7. Conclusion

We presented a modular text-guided multi-object tracking pipeline that combines open-vocabulary detection, classical association, and a prompt-consistency module to enforce semantic alignment with natural language descriptions. Our approach enables selective, mission-driven tracking without retraining and maintains tracking performance comparable to class-based baselines on standard benchmarks.

In addition to conventional MOT evaluation, we introduced a controlled CARLA-based prompt-tracking benchmark and a set of prompt-compliance metrics to measure semantic correctness with respect to natural language prompts. Future work will focus on improving robustness to appearance variation and illumination changes, exploring learned attribute representations to replace hand-crafted verification modules, and extending the framework toward more complex referring expressions and spatio-temporal reasoning in dynamic environments.

7.1. Limitations

While our approach enables prompt-driven and attribute-aware tracking without retraining, it has several limitations. First, CLIP-based similarity scores are not calibrated for fine-grained attributes such as color, which motivates our use of explicit HSV-based verification but may still fail under extreme illumination conditions or strong specular reflections. Second, the proposed pipeline relies on the quality of the open-vocabulary detector; persistent detector failures cannot be recovered through association alone. Third, the additional semantic checks introduce modest computational overhead compared to purely IoU-based trackers. Finally, abstract or ambiguous prompts (e.g., "the important car") are not supported and require explicit visual or spatial attributes.

Acknowledgment

This material is based upon work sponsored by the Defense Advanced Research Projects Agency (DARPA) and AFRL. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the DARPA or AFRL.

References

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019. 5

[2] Patrick Dendorfer, Aljoša Ošep, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, Stefan Roth, and Laura Leal-Taixé. Motchallenge: A benchmark for single-camera multiple target tracking. *arXiv preprint arXiv:2010.07548*, 2020. 2, 5

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 4

[4] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 370–386, 2018. 5

[5] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 2, 5

[6] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8844–8854, 2022. 3

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[8] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 2

[9] Dongming Wu, Wencheng Han, Tiancai Wang, Xingping Dong, Xiangyu Zhang, and Jianbing Shen. Referring multi-object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14633–14642, 2023. 2

[10] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829, 2024. 2

[11] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16442–16453, 2022. 2

[12] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*, pages 1–21. Springer, 2022. 1