GOVTrack: Towards Generative Open-Vocabulary Multi-Object Tracking
Supplementary Material

Zekun Qian[1] Ruize Han[2†] Zhixiang Wang[1] Liang Wan[1] Wei Feng[1]
[1]College of Intelligence and Computing, Tianjin University
[2]Faculty of Computer Science and Artificial Intelligence,
Shenzhen University of Advanced Technology
{clarkqian, zhixiang_wang, lwan, wfeng}@tju.edu.cn, hanruize@suat-sz.edu.cn

1. In-depth Analysis

1.1. Dataset Comparison

We further provide more statistics of GOVTrackB to show its data distribution and characteristics in Table 1. We can see that regardless of the total number of videos, frames, tracks, and bounding boxes, GOVTrackB is larger than the other two datasets. The "Density" metric means the object count in each frame, where GOVTrackB exceeds the other two datasets significantly. Additionally, for the average number of tracks per video, and the average number of boxes per video or track, GOVTrackB is also superior. These comparisons fully reflect the enrichment of GOVTrackB, which is beneficial for the evaluation of the generative open-vocabulary-related task and tracking tasks.

Table 1. Comparison of different datasets on various factors.

Datasets	# Vid	# Frm	# Trk	# BBox	Density	Trk/Vid	BBox/Vid	BBox/Trk
OVTAO-val	998	36K	5K	113K	1~11	5.5	114.2	20.6
OVTAO-burst	1,419	52K	8K	167K	1~11	5.6	117.5	21.0
GOVTrackB	1,635	55K	10K	266K	1~41	6.4	162.7	25.4

1.2. Reliability of the mgReA Metric

The calculation of the proposed mgReA metric involves the CLIP model for class matching. To assess its reliability, we compared the mgReA values for the generated classes obtained through our CLIP-matched method with those derived from manual annotations. Specifically, we randomly selected 40 videos from our GOVTrackB dataset and extracted detection bounding boxes based on an IoU threshold greater than 0.5 with the ground truth, resulting in a total of 2,350 detection samples. We then compared the mgReA metrics produced by both methods. As shown in the Table 2, the mgReA value for the base class was 26.65% using the CLIP matching method, while it was 25.05% for manual annotation. For the novel class, the mgReA scores were 23.30% for CLIP and 24.65% for annotation. The differences between these results were minimal, with the mgReA of the base class differing by only 1.60% and the novel class by 1.35%. This strongly supports the reliability and utility of the CLIP-based mgReA metric. It is important to note that the selected 2,350 manually annotated detection samples accounted for approximately 1% of the total bounding boxes in our dataset. Given the considerable labor involved, the use of CLIP-based mgReA metric in our study is deemed reasonable, despite the small differences observed.

Table 2. Results of mgReA with different matching methods.

	Base mgReA	Novel mgReA
CLIP matching	26.65	23.30
Annotator matching	25.05 (Δ1.60)	24.65 (Δ1.35)

1.3. Dataset Comprehensiveness

As discussed above, both OVTAO and the proposed GOVTrackB use the videos in the TAO dataset. We select the overlapped videos in OVTAO and GOVTrackB, and apply the public OVTrack [5] method on them for comparison. As shown in Table 3, we find that although the overlapped videos included in GOVTrackB represent only 40.3% of the original OVTAO-val and 41.2% of OVTAO-burst, the experimental results show negligible differences. The evaluated results on both base and novel categories diverge by no more than 0.6% for TETA, TRETA, when compared to the original OVTAO dataset. This comparison demonstrates that the portion of OVTAO dataset included in GOVTrackB *is highly representative, containing the data distribution diversity of the original OVTAO dataset*. Besides these videos, GOVTrackB also includes the data from LV-VIS. The richness of categories and quantity of samples have been significantly expanded in terms of principles **P1** and **P2**, making GOVTrackB highly effective and comprehensive for the GOVMOT.

†Corresponding author.

Table 3. Comparison results on datasets extracted from OVTAO in GOVTrackB (%).

Dataset	# Video	Base Class						Novel Class					
		TETA	LocA	AssocA	ClsA	mgReA	TRETA	TETA	LocA	AssocA	ClsA	mgReA	TRETA
OVTAO-val	998 (100%)	35.5	49.3	36.9	20.2	29.2	38.5	28.0	48.8	33.6	1.5	9.7	30.7
$\text{GOVTrackB}_{\text{OVTAO-val}}$	402 (40.3%)	36.1 (Δ0.6)	50.2	37.8	20.4	29.3	39.1 (Δ0.6)	27.8 (Δ0.2)	46.4	35.7	1.2	8.3	30.1 (Δ0.6)
OVTAO-burst	1419 (100%)	32.0	45.6	33.5	16.9	24.1	34.4	24.4	42.3	29.1	1.8	6.1	25.8
$\text{GOVTrackB}_{\text{OVTAO-burst}}$	585 (41.2%)	32.1 (Δ0.1)	45.5	34.4	16.4	24.0	34.6 (Δ0.2)	25.0 (Δ0.6)	43.1	29.7	2.3	6.4	26.4 (Δ0.6)

1.4. Visualization Analysis

The Fig. 1 presents some visualization results of GOVTracker, in which bounding boxes with the same color indicate the same track ID, text boxes with a black background display the generated category names (prediction), while the text boxes with a green background show the labels obtained using CLIP for evaluation, and the text boxes with a brown background indicate the ground-truth labels in the dataset. We can see that GOVTracker encapsulates a rich understanding of object categories. For instance, in the first row, GOVTracker is not only able to identify the object as a "child", but also recognizes it as a "girl", thereby providing a more comprehensive description of the target. Importantly, this is achieved without the need for any pre-specified category restrictions. In the second row of results, the generated output includes the prediction "grizzly bear", even more specific than the ground truth "bear". The third row demonstrates the effectiveness of the proposed mgReA in Section 3.5 of the main paper. We can observe that for tracking a specific subclass "dalmatian" of "dog", GOVTracker effectively describes the target's characteristics, such as "black and white dog". It can also predict its super-category "dog" and accurately identify the subclass "dalmatian" in certain frames. When the target is recognized as "dog", the multi-granularity metric mgReA traces back to the expanded label "dog" from the ground-truth label "dalmatian", effectively addressing the misalignment between the generated results and ground truth labels.

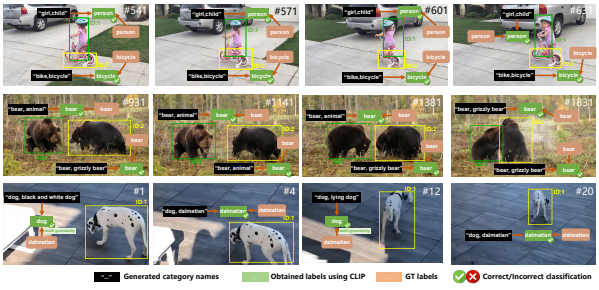

Figure 1. Illustration of the qualitative results of GOVTracker.

Qualitative comparison on Internet data. We also use a video from the Internet (involving a "chickadee") not included in the dataset for testing, to investigate the effectiveness of different methods in real-world applications. As shown in Fig. 2, OVTrack requires the input configuration of not only the video but also the predefined category list. Following the original setting in OVTrack, the category list contains 1,203 classes (base + novel) predefined in LVIS. As shown in the first row, OVTrack (with input ① and ②) only provides the results of the broad labels like "bird" (appearing in the category list), without additional fine-grained categories. This is because the category "chickadee" is not included in the given category list. As shown in the second row, we further add the category "chickadee" to the category list for OVTrack. This way, OVTrack (with input ①, ②, and ③) can recognize the object (in some cases). While additional category inputs can improve classification accuracy, the results remain unstable during tracking. More importantly, for the objects outside the predefined categories, additional manually curated categories are essential, which, however, is troublesome and impractical. Conversely, GOVTracker simplifies the process by using only the video itself as input, without requiring any other information. As presented in the last row, GOVTracker's results (with only input ①) demonstrate the capability of its generative model to produce detailed results, identifying not only general categories such as "bird" but also specific subcategories like "chickadee", and even detailed descriptions like "round bird", a result previously unattainable in any MOT system. These findings highlight that GOVTracker requires simple input while delivering more comprehensive and varied outputs compared to OVTrack. This case intuitively verifies the advantages of GOVMOT over OVMOT.

Figure 2. Qualitative comparison of the methods for OVMOT and GOVMOT.

Failure case analysis. We also present a representative failure case in Fig. 3. This failure example includes an ID switch caused by target appearance differences, and the misclassification of similar categories. Specifically, the target with ID 1 at frame # 541 is failed to be tracked at # 631. This

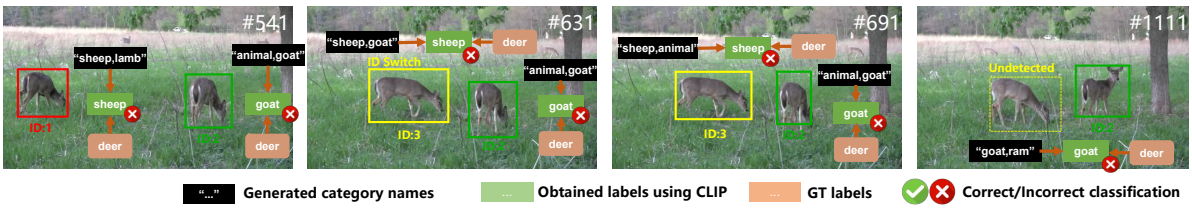

Figure 3. Qualitative analysis of a failure case.

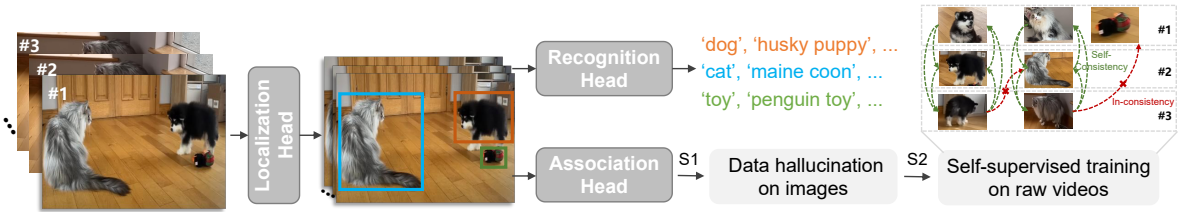

Figure 4. Pipeline of the proposed baseline method GOVTracker.

may be caused by the irregular motion and shape variation during this period. For recognition, the "deer" appearing in this video is incorrectly recognized as the "sheep". This is because they look very similar, and the viewing angle makes the video not contain enough discriminative characteristics. From this point, while acknowledging the practical value of GOVTracker, we also recognize the ample room for improvement in addressing the various challenges in the GOVMOT problem.

2. Implementation Details

The Fig. 4 illustrates the framework of our pipeline, which adheres to the structure of most tracking-by-detection frameworks. It consists of a localization head, a recognition head, and an association head.

In the localization module, we use Swin Transformer [6] as the backbone for visual encoding. Following [8], the Deformable DETR architecture has 6 encoder layers and 6 decoder layers. The number of object queries is set to 300.

In the recognition module, we use the FlanT5-base [1] and initialize it with its pre-trained weights. Flan T5-Base is a Transformer model that includes both encoder and decoder structures. We use visual features of the candidate objects obtained from Deformable DETR are mapped to the input space of the generative model through a projection layer, and then processed by a generative encoder and decoder, both composed of self-attention layers and feedforward neural networks. The encoder's output interacts with the decoder through the cross-attention layers. Then the decoder's output is passed through a softmax layer to predict the corresponding word, while the prediction of the previous word is used as input for training the next word prediction. We select this network due to its parameter count of only 250M, which effectively reduces computational complexity. Additionally, its performance is significantly enhanced through efficient multi-task instruction fine-tuning. The beam size is set to 2, *i.e.*, the output number of categories of each object is 2.

In the association module, the data hallucination strategy uses the same setting as OVTrack [5] with a learning rate of 2×10^{-3}. For the consistency-aware focal loss, we set β_1, β_2, and α as 0.7, 0.3, 0.9, respectively. In the self-supervised strategy, we use the TAO training dataset [2] without any annotation for training. We select a continuous sequence of 24 frames as a batch, grouping them using C_N^2 and C_N^3 combinations to enhance the data scale. Then, we employ the cyclic-consistency loss proposed in the work [4] for self-supervised training with a learning rate of 2×10^{-4}.

In the training stage, we first train the localization and recognition modules together for 20 epochs. The learning rate of the backbone and detection head is set to 2×10^{-4}, and the language model is set to 3×10^{-4}. Then we train the association head using a hallucination strategy with consistency-aware focal loss for 6 epochs and self-supervise the association head for 14 epochs. In the inference stage, we associate the historical tracks to the objects detected in the current frame using the appearance feature similarity obtained from the trained association head. Following the methodology in [5], we compute the similarity between historical tracks and detected objects using both bi-directional softmax [3] and cosine similarity metrics. In line with traditional MOT methods, we assign an object to a track if the similarity score surpasses a matching threshold. If an object doesn't match any existing track, a new track is initiated if its detection confidence score from the classification head exceeds a threshold; otherwise, it is ignored.

We use PyTorch as the implementation framework and conduct the experiments on a server with 8 RTX 3090 GPUs. The optimizer used in our method is the AdamW[7] with a

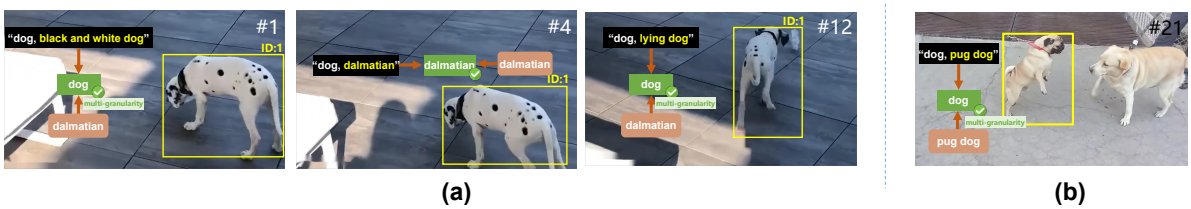

Figure 5. Qualitative analysis of limitations.

weight decay of 0.05.

3. Discussion and Limitation

3.1. Experiment Insights

From the comparison results in Table 1 of the main paper, we can observe that the performances of all the methods are generally poor, especially for the recognition task. This reflects *the challenges of GOVTrackB and also the GOVMOT problem*, which have great space for improvement. Although the GOVMOT task is more challenging and does not use a class list as input, it achieves performance comparable to that of the latest OVMOT trackers while setting even better performance benchmarks. This indicates that GOVMOT is practically feasible and more versatile, as it eliminates the need for a class list and enables effective tracking with multi-granularity perception of every real-world object. This demonstrates that *the proposed more practical GOVMOT task is very promising*.

3.2. Limitation and Future Work

We also discuss the limitations of the proposed GOVTracker. First, we find that although the association mechanism can effectively track targets in many cases, it generates different category nouns for the same track ID, resulting in a lack of consistency and coherence in the generated object categories. As shown in Fig. 5(a), the generated categories for this dog include "black and white dog", "dalmatian" and "lying dog" in the same track, which are not consistent. Such changes in the predicted object categories during tracking make it difficult for the tracker to accurately maintain a stable category. To address this problem, in future work, we plan to study the mutual collaboration mechanism between tracking and recognition in GOVMOT, *i.e.*, using continuous tracking to help the consistent category prediction, as well as making use of the category estimation results to assist object tracking. Second, as discussed in the "evaluation metrics" part of Section 3.5 in the main paper, due to the diversity of generated target category nouns, we use the CLIP to match these nouns with the most similar label from the GT category set. The rationale behind this approach is to give a more appropriate evaluation, through comprehensive semantics matching with the help of pre-trained multi-modal models. This metric is reasonable. However, it may not be the optimal solution. We find that sometimes the correct category noun is present among the generated nouns, but the CLIP-based matching still classifies it into a more general category. As shown in Fig. 5 (b), we can see that although the generated target category by our method includes "pug dog," the final label matched by CLIP used for evaluation is "dog." Since the ground-truth label is "pug dog', if we directly use this matching label "dog" for judging, the evaluation result will be "false". But, fortunately in this case, if we use the multi-granularity recognition (mgReg) metric, the evaluation result will become "true". Even though, we still think that the overall evaluation system for GOVMOT has space for improvement.

We have discussed some issues that were not addressed in this work. As the first GOVMOT benchmark, we aim to promote the expansion of tracking tasks with various constraints in existing works to more general and practical tracking. This will significantly advance the progress of the tracking community. We hope that with our efforts or those of other researchers in the future, these limitations can be partly alleviated or effectively resolved.

4. Access to Dataset and Benchmark

The dataset, related benchmark code, and Croissant metadata in this paper can be found at: https://github.com/zekunqian/GOVTrackB.

References

[1] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research (JMLR)*, 25(70):1–53, 2024.

[2] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. TAO: A large-scale benchmark for tracking any object. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 436–454, 2020.

[3] Tobias Fischer, Thomas E Huang, Jiangmiao Pang, Linlu Qiu, Haofeng Chen, Trevor Darrell, and Fisher Yu. QDTrack: Quasi-dense similarity learning for appearance-only multiple object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023.

[4] Yiyang Gan, Ruize Han, Liqiang Yin, Wei Feng, and Song Wang. Self-supervised multi-view multi-human association and tracking. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 282–290, 2021.

[5] Siyuan Li, Tobias Fischer, Lei Ke, Henghui Ding, Martin Danelljan, and Fisher Yu. OVTrack: Open-vocabulary multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5567–5577, 2023.

[6] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

[8] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*, 2021.