DTLLM-VLT: Diverse Text Generation for Visual Language Tracking Based on LLM

Xuchen Li[1] Xiaokun Feng[1,2] Shiyu Hu[1,2] Meiqi Wu[3]
Dailing Zhang[1,2] Jing Zhang[1] Kaiqi Huang[1,2,4]
[1]CRISE, Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]School of Computer Science and Technology, University of Chinese Academy of Sciences
[4]CAS Center for Excellence in Brain Science and Intelligence Technology
{lixuchen2024, fengxiaokun2022, hushiyu2019, zhangdailing2023, jing_zhang, kqhuang}@ia.ac.cn,
wumeiqi18@mails.ucas.ac.cn

Abstract

Visual Language Tracking (VLT) enhances single object tracking (SOT) by integrating natural language descriptions from a video, for the precise tracking of a specified object. By leveraging high-level semantic information, VLT guides object tracking, alleviating the constraints associated with relying on a visual modality. Nevertheless, most VLT benchmarks are annotated in a single granularity and lack a coherent semantic framework to provide scientific guidance. Moreover, coordinating human annotators for high-quality annotations is laborious and time-consuming. To address these challenges, we introduce ***DTLLM-VLT****, which automatically generates extensive and multi-granularity text to enhance environmental diversity. (1) DTLLM-VLT generates scientific and multi-granularity text descriptions using a cohesive prompt framework. Its succinct and highly adaptable design allows seamless integration into various visual tracking benchmarks. (2) We select three prominent benchmarks to deploy our approach: short-term tracking, long-term tracking, and global instance tracking. We offer four granularity combinations for these benchmarks, considering the extent and density of semantic information, thereby showcasing the practicality and versatility of DTLLM-VLT. (3) We conduct comparative experiments on VLT benchmarks with different text granularities, evaluating and analyzing the impact of diverse text on tracking performance. Conclusionally, this work leverages LLM to provide multi-granularity semantic information for VLT task from efficient and diverse perspectives, enabling fine-grained evaluation of multi-modal trackers. In the future, we believe this work can be extended to more datasets to support vision datasets understanding.*

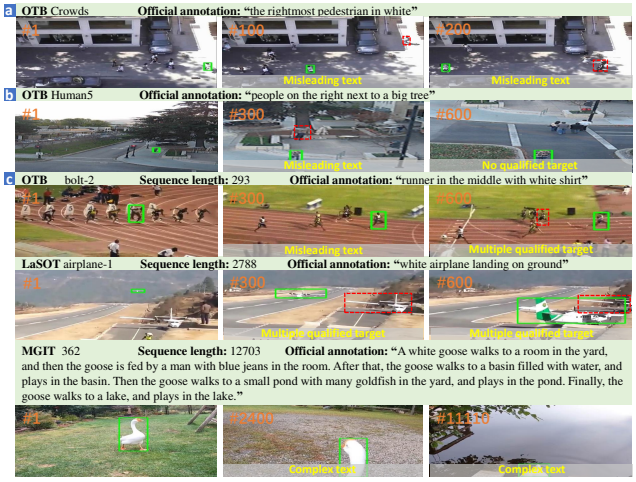

Figure 1. Examples of video content and semantic descriptions on OTB99_Lang [19], LaSOT [3], and MGIT [9] benchmarks. The green bounding box (BBox) indicates ground truth, while the red dashed BBox indicates other objects that satisfy the semantic description. (a) and (b) are short sequences in OTB99_Lang with simple narrative content. Besides, their semantic annotations mainly describe the first frame, which may misguide the algorithm. (c) Comparison of different text annotations, video length, and content on three benchmarks. The VLT environment is complex, variable and most of them suffer from issues of inconsistent text styles and single annotation granularity.

1. Introduction

Single object tracking (SOT) is a crucial computer vision task focused on tracking a moving object within a video sequence. Researchers have consistently observed the limited performance of most trackers in long videos with more complex video content. Moreover, relying solely on a vi-

sual modality greatly constrains the versatility of such systems. Consequently, several studies have begun providing semantic annotations for the SOT task, leading to the emergence of the visual language tracking (VLT) task. The proposal of VLT task helps the research of SOT to be more human-like and broaden its application prospects. Natural language, in contrast to bounding boxes (BBox), provides a more user-friendly and intuitive way of describing objects, allowing for precise descriptions ranging from spatial locations to high-level semantic details to improve tracking performance. When defining the VLT task, researchers incorporate text annotations from two main viewpoints:

(1) *Short text annotation.* Representative VLT benchmarks such as OTB99_Lang [19], TNL2K [27], and LaSOT [3, 4] primarily employ short text. This concise style of description is clear and uncomplicated, facilitating the learning and comprehension of VLT trackers. The utilization of short text offers the benefit of simplicity and enhanced comprehension for VLT trackers. However, these methods are prone to imprecise semantic descriptions and potential ambiguities. As illustrated in Fig. 1 (a) and (b), the description only captures the state of the object at the sequence beginning. As the object moves, the positional constraint in the semantic information becomes misleading. The reason lies in the benchmark focus primarily on the initial state of the object, neglecting changes in the object's motion throughout the video. Consequently, semantic descriptions may become restrictive later in the sequence.

(2) *Long text annotation.* MGIT [9] adopts a multi-granular semantic annotation strategy from the perspective of more precise semantic descriptions, providing a way to annotate complex spatio-temporal causal relationships in long videos. Compared to other benchmarks, this style exhibits two characteristics: longer text and periodic updates, evolving from simple to dense, detailed descriptions. However, this approach faces challenges like time-intensive text annotations and the need for algorithms with robust text processing and multi-modal alignment capabilities to effectively utilize the information. As shown in Fig. 1 (c), the text in MGIT is overly long and complex. Clearly, although the motivation of these works is to extend SOT task to multi-modal one to enhance tracking performance, the disparate styles and singular granularity across most studies not only hinder algorithms from achieving the desired outcomes but also escalate the complexity of research on VLT task.

In summary, diverse motivations in existing research result in varying approaches to integrating textual information. In Fig. 1 (c), the three prominent benchmarks differ in sequence length, text style, and annotation granularity. Imposing a single standard mechanism for VLT research appears impractical, given the inherent flexibility and variability in human comprehension and processing of multi-modal information. Humans can adeptly leverage various types of multi-modal information. Rather than enforcing a rigid task format, ***optimal design should furnish algorithms with comprehensive environmental data to explore their capabilities and limitations***.

By offering diverse text descriptions of the environment—encompassing short, long, sparse, and dense formats—and evaluating algorithm performance across these descriptions, we can effectively discern the strengths and weaknesses of existing methods under different semantic granularities, thereby guiding the enhancement of multi-modal algorithms. What excites us is that the Large Language Model (LLM) can facilitate the achievement of this goal. By seamlessly integrating the LLM into the text generation process, we can offer a varied multi-modal environment conducive to VLT research.

Our work focuses on the aforementioned motivations and designs DTLLM-VLT to achieve diverse text generation for tracking datasets. Specifically, we combine text length and generation density to form four granularities with a uniform style. Based on this, we select MMTrack [34], a state-of-the-art (SOTA) VLT tracker, for experimental analysis to verify the impact of diverse texts on algorithm performance. The experimental results not only demonstrate that this diversified environment can assist in fine-grained evaluation and analysis of algorithm capabilities but also suggest the possibility of further enhancing the multi-modal learning capabilities of algorithms using generated data in the future.

The contributions of this paper can be summarized in the following three aspects:

- We develop DTLLM-VLT, a model based on LLM, aimed at efficiently generating high-quality scientific text for tracking datasets at scale. DTLLM-VLT can seamlessly apply to various tracking tasks.
- We generate diverse text for three prominent VLT benchmarks, addressing four levels of granularity. This approach overcomes the limitations of previous benchmarks, which focused on a single granularity and lacked a unified semantic framework.
- We conduct an experimental analysis to evaluate the impact of diverse texts on algorithm performance. The results highlight the benefits of a diversified environment and indicate the potential for enhancing multi-modal learning through generated text data.

2. Related Work

2.1. Single Object Tracking Benchmark

The SOT task involves initializing and tracking a specific object within a video sequence. It begins by identifying the object through its BBox in the first frame and then proceeds to locate and follow the object across subsequent frames. Since 2013, several benchmarks such as OTB [28, 29] and

Table 1. Summary of current popular tracking benchmarks and Comparison number of language description between official and our generated text. *Italics* indicate automatic generation. We provide far more diverse semantic information than the original annotations for representative environments.

Dataset	Number of Videos		Number of Language Description				
	Train	Evaluation	Official	*Dense Concise*	*Dense Detailed*	*Initial Concise*	*Initial Detailed*
OTB99_Lang [19]	51	48	99	*596*	*596*	*99*	*99*
LaSOT [3]	1,120	280	1,400	*35.2K*	*35.2K*	*1,400*	*1,400*
TNL2K [27]	1,300	700	2,000	*12.4K*	*12.4K*	*2,000*	*2,000*
MGIT[1][9]	105	45	1,753	*16.1K*	*16.1K*	*120*	*120*

[1] As the ground truth of the MGIT [9] test set is not open-sourced, we only generated text for 120 video of the training and validation sets.

VOT [14, 15] have been introduced, providing standardized datasets and scientific evaluation mechanisms to support SOT research. However, with the advancements in deep learning techniques, these short-term and small-scale benchmarks have faced challenges in adequately accommodating data-driven trackers. Consequently, researchers have started designing larger-scale datasets such as GOT-10k [12] and TrackingNet [20]. Additionally, efforts have been made to gather data featuring long videos, leading to the creation of long-term tracking benchmarks like OxUvA [24] and VOT_LT [16, 17]. Some work has also focused on SOT in drone scenarios, such as BioDrone [33], a vision benchmark for SOT based on bionic drones. Recently, researchers have acknowledged that traditional approaches to both short-term and long-term tracking are based on the premise of constant movement, a factor that restricts testing to situations involving a single camera view and a static scene. To expand beyond these limitations, they have introduced the global instance tracking task along with a novel benchmark called VideoCube [10], which enables the tracking of arbitrary moving objects in various types of videos. To scientifically evaluate the performance of trackers under different challenging factors, researchers have introduced SOTVerse [11], a user-defined space for SOT task.

2.2. Visual Language Tracking Benchmark

While visual benchmarks have undergone significant evolution over the past decades, benchmarks integrating visual and semantic information, known as VLT benchmarks, have only recently gained traction. OTB99_Lang [19] stands out as the first VLT benchmark, enhancing sequences from the OTB100 [29] benchmark with additional natural language descriptions. However, the limited scale of the dataset has hindered the widespread adoption of the VLT task. Subsequently, the release of LaSOT [3, 4], a long-term tracking benchmark with natural language annotations, marked a significant development. Concurrently, researchers introduced the TNL2K [27] benchmark in the same year, aiming to enhance object tracking flexibility and accuracy through text descriptions. Following these efforts, researchers proposed a new multi-modal benchmark named MGIT [9], which fully represents the complex spatio-temporal and causal relationships present in long narrative content through a multi-granular annotation strategy. These three benchmarks have enriched the pool of available data and facilitated the development of various VLT trackers.

2.3. Algorithms for Visual Language Tracking

VLT emerges as a burgeoning multi-modal task aiming to achieve tracking by leveraging both a language description and an initial template patch. Following the principle of similarity-matching, most existing VLT methods [5, 6, 8, 18, 25, 26, 32] utilize language descriptions and template patches as references to identify the most similar object in the search frame. Among these methods, SNLT [7] presents an adaptable language-based region proposal network that improves tracking accuracy by employing a dynamic aggregation mechanism. Meanwhile, MMTrack [34] introduces a streamlined and effective tracking method, treating the VLT task as a sequence of token generation. However, these methods often fail to capture the dynamic properties of the object, which becomes a critical issue for robust tracking when the object's appearance undergoes significant changes. To overcome this shortcoming, some VLT trackers have begun to integrate temporal data to establish a more dynamic reference. For instance, GTI [30] and AdaSwitcher [27] identify object by merging tracking and localization outcomes at every time interval. JointNLT [35] also takes a step towards this by including temporal information as queries during the prediction phase.

Most benchmarks for VLT provide only one natural language description per video. Additionally, the existing benchmarks suffer from inconsistent text annotation styles, leading to varied mechanisms for incorporating text information. These discrepancies hinder algorithm evaluation and comprehension of video content. Moreover, these works all provide semantic information in the form of manually annotated data, which is a time-consuming and labor-intensive process.

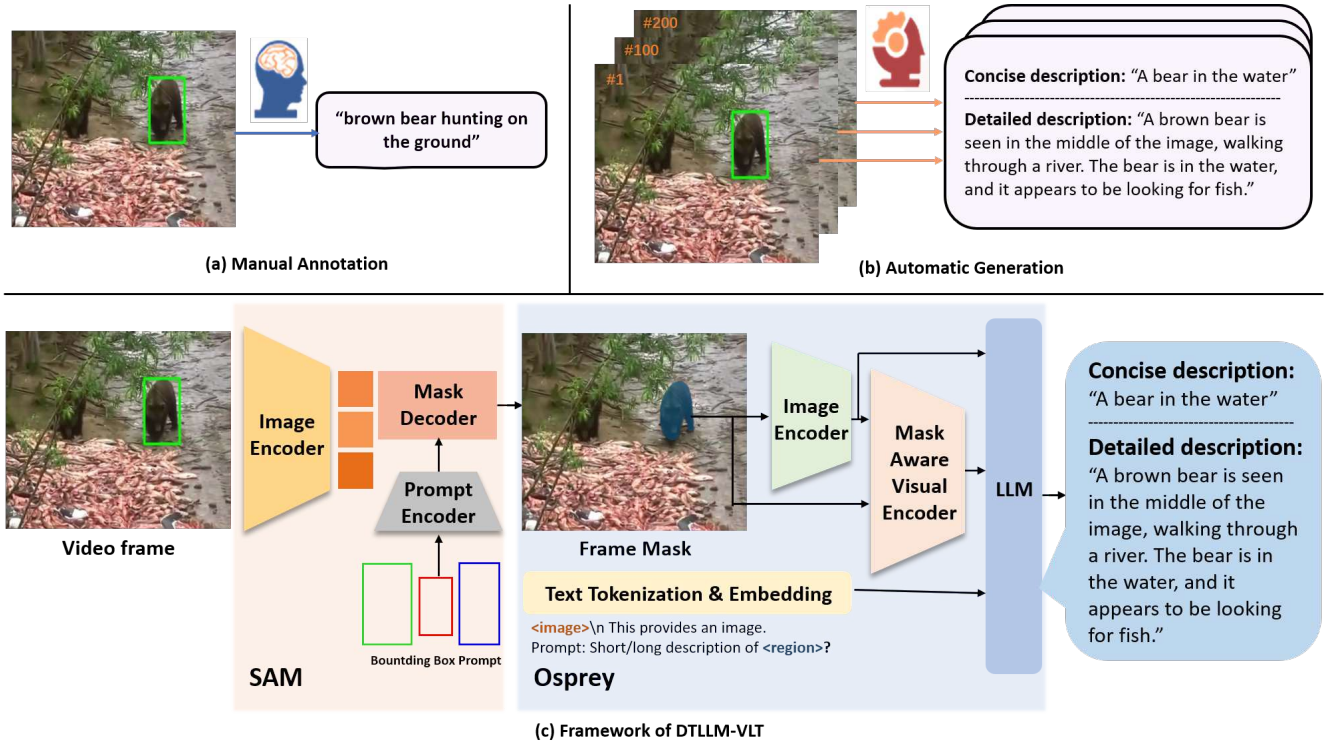

Figure 2. Comparison of Manual Annotation and Automatic Generation and Framework of DTLLM-VLT. (a) Manual annotation relies on human labor, only provides one text annotation for each video segment, and cannot guarantee a uniform style. The cost of large-scale annotation is too high. (b) Automatic Generation can generate diverse text on a large-scale in a unified style. (c) The DTLLM-VLT can provide dense concise/detailed text generation based on given video frames and BBox of object.

3. Text Generation by LLM

To provide diverse text generation for VLT datasets under a unified prompt framework and provide algorithms with more scientific text for evaluation and understanding video content, we implement DTLLM-VLT to offer large-scale automatic diverse generated text.

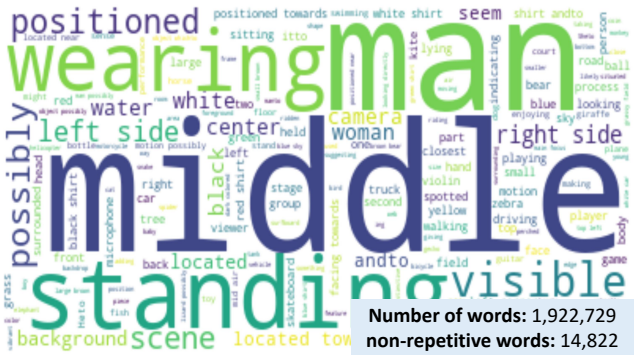

Figure 3. The word cloud of semantic descriptions and word count statistics.

3.1. Generation Strategy

The volume and linguistic annotations of the VLT dataset determine the quality and generality of learned visual language representations. Table 1 illustrates that the dataset comprises only 3,649 videos, specifically 1,400 from LaSOT [3], 2,000 from TNL2K [27], 99 from OTB99_Lang [19], and 150 from MGIT [9], which are used for training and testing. These videos are accompanied by 5,252 official text descriptions. However, this amount of data is deemed insufficient for algorithms to effectively learn.

These official annotation suffers from inconsistency in style, and are only able to describe short-term changes for the object. The varying annotation styles of the text descriptions make it difficult for trackers to learn general visual language information, resulting in a significant performance drop when inferring on new videos with non-official annotations or different language description styles. Moreover, inaccurate text descriptions hinder object tracking, turning natural language annotations into a hindrance rather than a support.

To enhance the accuracy and generality, we propose DTLLM-VLT, which generates text in a consistent style for four datasets, establishing a robust foundation for VLT. This

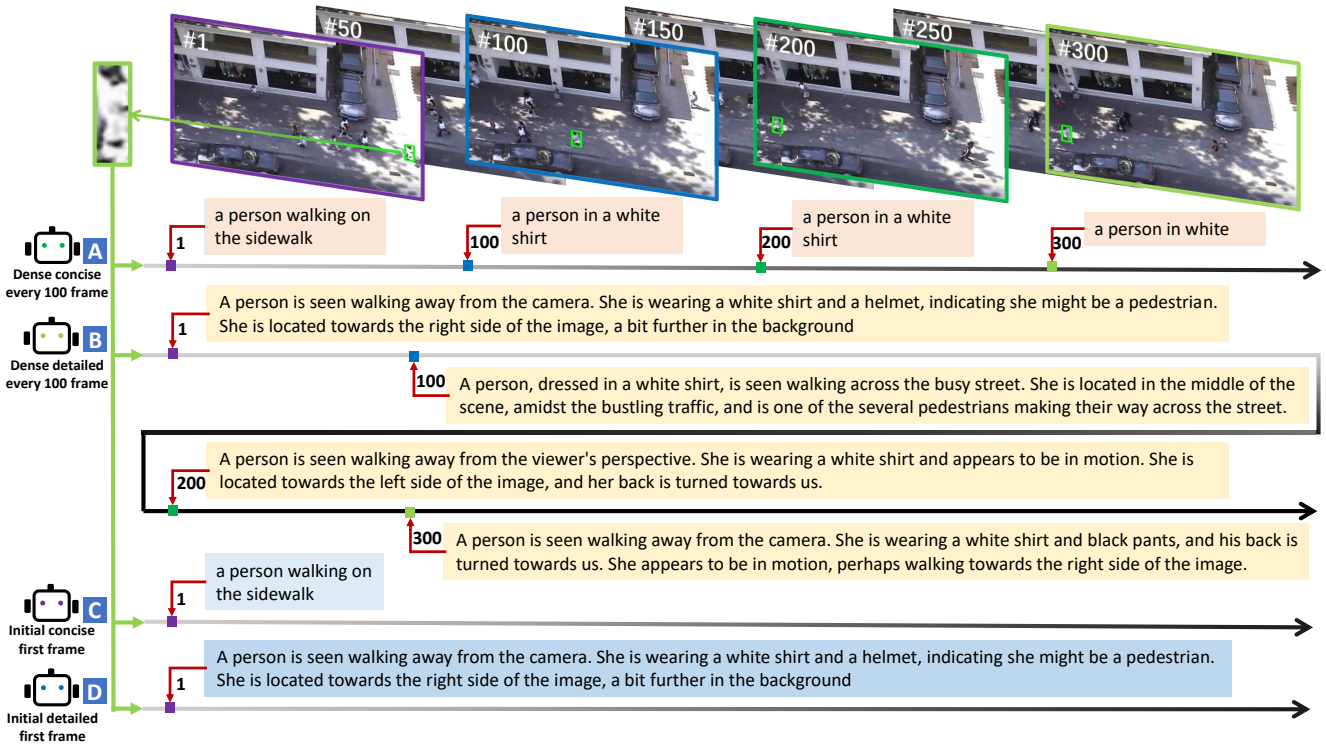

Figure 4. Examples of the four types of generated text. We provide four different natural language descriptions for each video. The object to be tracked is determined in the first frame and does not change throughout the video sequence.

generation approach can be expanded to additional VLT datasets and even applied to text generation in SOT datasets.

Initial and dense text descriptions. Following the text annotations method in OTB99_Lang [19] and TNL2K [27], we generate text for the initial frame of each video. Additionally, given that 4 seconds marks the threshold between human instant memory and short-term memory [1, 21, 22], we consider the worst situation and infer that the algorithm lacks an efficient memory system. Consequently, at 25 FPS, equating to every 100 frames in 4 seconds, we supply the algorithm with relevant generated text. We posit that this update frequency optimally sustains the memory state of algorithm and enhances tracking performance.

Concise and detailed text descriptions. For the algorithm, if the BBox already sufficiently describes the temporal and spatial changes of the object, the text descriptions should focus on providing essential semantic details like the category and positions of the object. In cases where the BBox lacks sufficient information for effective learning by the tracker, more elaborate texts are necessary to compensate for the missing temporal and spatial relationships. Consequently, we generate two types of text descriptions: concise and detailed. As illustrated in Fig. 2, the concise text conveys essential information about the object, such as its category (*bear*) and position (*in the water*), while the detailed text includes additional spatio-temporal details like color, relative position, and actions.

3.2. DTLLM-VLT

The traditional VLT datasets rely on manual text annotations, as shown in Fig. 2 (a), providing a corresponding natural language description for each video. This method incurs high annotation costs, lacks uniformity in style, involves a single annotation granularity, and cannot be used for large-scale data annotation. To address these issues, we design DTLLM-VLT based on SAM [13] and Osprey [31], which can provide large-scale and diverse text generation like Fig. 2 (b).

The framework of the DTLLM-VLT is illustrated in Fig. 2 (c). Input video frames and corresponding object BBox, SAM [13] utilizes image encoder, prompt encoder, and mask decoder to obtain masks of the corresponding object and then input the video frames and mask into Osprey [31]. Osprey encodes the images and masks, combines with preset prompts, and generates concise and detailed descriptions of the corresponding object through LLM [2, 23]. Through this approach, we can generate large-scale, diverse granularities, and uniform style text for SOT and VLT datasets at very low costs.

3.3. Generation Analysis

Combining the aforementioned strategies, we offer four granularities of natural language descriptions for each video, namely initial concise description, initial detailed description, dense concise description, and dense detailed description, as illustrated in Fig. 4. Our goal is to incorporate multiple granularities of text to enrich the environment for algorithm to learn and evaluate, while also providing guidance for algorithm design and model optimization.

Leveraging the DTLLM-VLT, we generate text descriptions comprising 7,238 initial descriptions (3,619 concise and 3,619 detailed descriptions each) and 128.4K dense descriptions (64.3K concise and 64.3K detailed descriptions each). Our dense texts are 24.4 times the quantity of the official annotations. Further details regarding the number of semantic descriptions are presented in Table 1. The semantic descriptions contain 1.9M words with 14.8K non-repetitive words. The vocabulary is rich, allowing for a comprehensive description of changes in the object during the tracking process. Word cloud and more detailed analyses have been illustrated in Fig. 3.

3.4. Speed and Memory Usage

We generate diverse text for visual language tracking datasets on RTX-3090 GPUs, with approximately 16GB of VRAM usage. It takes about 2 seconds to generate a text entry for each frame.

Compared to manual annotation, DTLLM-VLT can generate texts of various granularities for large-scale tracking datasets in a short period of time. And it can seamlessly apply to various tracking tasks.

4. Experimental Results

4.1. Datasets and Evaluation Methods

Datasets. We selected three representative datasets, OTB99_Lang [19], LaSOT [3], and MGIT [9], for evaluating short-term tracking, long-term tracking, and global instance tracking task. OTB99_Lang [19] and LaSOT [3] are expanded from the traditional SOT benchmark by adding language annotations. OTB99_Lang serves as a representative dataset for short-term tracking task, providing a text description for the initial frame of each video sequence. LaSOT is a representative dataset for long-term tracking task. Its text annotations only describe the appearance of the target, omitting relative positions. MGIT [9] is a novel large-scale benchmark specifically tailored for the global instance tracking task. Text annotations of each sequence contain complex spatio-temporal causal relationships with a multi-granular annotation strategy.

Evaluation Methods. As shown in Fig. 4, we follow generation granularities to design various mechanisms. We select a SOTA visual language tracker, MMTrack [34] as a baseline model and evaluate it on three benchmarks (as shown in Table 2 and Table 3). Compared with other algorithms, MMTrack [34] does not impose restrictions on the length of the text and does not truncate excessively long text. Additionally, it unifies the VLT task as a form of token generation, which is more conducive to learning visual language information.

To fairly compare the tracking performance on three datasets, we directly use the officially provided weights to test with the official annotations, initial concise texts, initial detailed texts, dense concise texts, and dense detailed texts. We also retrain and test the model under the corresponding settings to evaluate Area Under the Curve (AUC), tracking precision (P), and normalized precision (P_{Norm}).

4.2. Tracking Results

We evaluate MMTrack [34] on three benchmarks, including OTB99_Lang [19], MGIT [9], and LaSOT [3] with five text granularities to evaluate the influence of diverse generated text on tracking performance. All experiments employ joint language and BBox initialization.

4.2.1 Testing Directly

We directly use the model provided by the official for testing, and the test results are as shown in Table 2.

Short-term tracking. In Table 2, when comparing results on OTB99_Lang [19], which only provides the text description of the initial frame and will interfere with the tracking of the object in the later stage, our initial concise text achieves gains of 1.6 %, 2.2 %, and 1.6 % in area under the curve, normalized precision, and precision score, respectively. At the same time, we find that dense concise text also helps improve tracking performance, for example, our generated text achieves improvements of 1.2 % in the area under the curve. We think that the short-term tracking datasets represented by OTB99_Lang [19], their BBox can effectively describe the temporal and spatial relationships in the visual modality. If only the text from the initial frame is used and cannot describe the temporal and spatial relationships of the object in the following frame, it will cause significant interference. The same problem arises in our detailed initial concise/dense text description testing. In this case, the text only needs to be as concise as possible to assist in improving tracking performance.

Long-term tracking. The official text annotation of LaSOT [3] only describes the appearance of the object, ignoring the relative position. Compared to OTB99_Lang [19], the text description of the object is more accurate. Compared with MGIT [9], there is no excessive interference from relative position information. It represents a balance between the two and is most in line with the current algorithm learning method. Therefore, the test performance

Table 2. Comparison with testing directly on three popular benchmarks: OTB99_Lang [19], MGIT [9], and LaSOT [3]. The best two results are highlighted in red and blue, respectively.

Method	OTB99_Lang [19]			MGIT [9]			LaSOT [3]		
	AUC	P_{Norm}	P	AUC	P_{Norm}	P	AUC	P_{Norm}	P
Official	69.0	82.0	89.5	73.5	77.2	54.3	69.9	82.2	75.7
Initial Concise	70.6	84.2	91.1	73.9	77.8	54.9	69.0	81.1	74.7
Initial Detailed	68.0	81.5	88.4	72.7	76.2	53.4	68.7	80.7	74.4
Dense Concise	70.2	84.0	90.8	74.2	77.9	55.0	69.1	81.3	74.8
Dense Detailed	68.6	82.4	89.4	72.9	76.6	53.5	69.0	81.1	74.7

Table 3. Comparison with retraining and testing respectively on three popular benchmarks: OTB99_Lang [19], MGIT [9], and LaSOT [3]. The best two results are highlighted in red and blue, respectively.

Method	OTB99_Lang [19]			MGIT [9]			LaSOT [3]		
	AUC	P_{Norm}	P	AUC	P_{Norm}	P	AUC	P_{Norm}	P
Official	69.0	82.0	89.5	73.5	77.2	54.3	69.9	82.2	75.7
Initial Concise	70.0	84.3	90.5	73.6	77.4	54.2	69.6	81.8	75.4
Initial Detailed	70.3	85.6	91.4	74.1	78.3	54.5	69.4	81.5	75.1
Dense Concise	71.3	86.0	92.5	74.0	77.6	54.2	69.5	81.6	75.3
Dense Detailed	69.8	84.8	90.6	74.4	78.5	54.6	69.8	82.1	75.6

with official annotation is the best. However, we believe that for long-term tracking, providing only a single sentence of text is not conducive to algorithm learning. And the spatial relationships of the object are crucial. When there are large-scale and diverse VLT datasets and better approaches to enhancing video understanding capabilities of algorithm, this situation observed in LaSOT [3] will soon change.

Global instance tracking. The same situation as OTB99_Lang [19] appeared on MGIT [9], that is, the performance is improved when tested under initial/dense concise text annotations. Particularly, dense concise annotation excels over the official text, surpassing it by 0.7 %, 0.7 %, and 0.7 % in area under the curve, normalized precision, and precision score, respectively. MGIT [9] provides high-quality, multi-granularity long texts containing complex temporal and spatial relationships. From the test results, we think that the handling of long texts and multi-modal alignment in the current algorithm requires improvement, as it fails to fully leverage temporal and spatial relationships. Therefore, concise text can actually help improve performance. However, temporal and spatial information are crucial for long-term tracking and global instance tracking. When the temporal-spatial information of the BBox cannot stably determine the object, detailed text is needed to provide additional high-level semantic information to identify the object.

Through direct testing and comparison of tracking performance under different texts, it has been observed that the variation in texts has a significant impact on tracking performance. The largest performance difference reached 2.2% in normalized precision on the OTB99_Lang dataset.

4.2.2 Retraining and Testing Respectively

As mentioned earlier, when the dataset text becomes denser and more accurate, it can compensate for BBox shortcomings. The algorithm gains additional knowledge through text updates, potentially improving performance. Therefore, we retrained and tested MMTrack [34] using varied generated texts, with tracking results shown in Table 3.

Short-term tracking. It can be seen that on the OTB99_Lang [19] benchmark, the testing results after retraining with dense concise text have shown further improvement. Compared with the official text, it gains 2.3 %, 4.0 %, and 3.0 % in area under the curve, normalized precision, and precision score, respectively. This indicates that providing dense concise text on short-term datasets can further improve tracking performance. It also reflects the capability of the current algorithm to achieve better tracking even when provided with more accurate text, without the need for matching learning methods. However, we believe that the current method of training algorithms to memorize high-frequency text for enhancing memory capabilities still needs improvement, the potential of text has not been fully exploited yet.

Long-term tracking. The results on the LaSOT [3] benchmark show that official annotations are still more advantageous for tracking. However, after retraining, the results on dense detailed text are only 0.1 % from the optimal

results, indicating an improvement in the algorithm's understanding of dense text compared to direct testing, but it is still unable to fully learn all temporal and spatial information.

Global instance tracking. The test results after retraining based on different texts show that the algorithm can improve its tracking ability on the MGIT [9] benchmark by learning from dense detailed text, which differs from the results of direct testing. For global instance tracking task, it is beneficial for tracking if the algorithm can learn more comprehensive temporal and spatial relationships.

Comparing the above results, we can draw the following insights:

(1) **The existing algorithm tends to learn and understand short text.** The results of direct testing show that concise text is more beneficial for performance improvement on the OTB99_Lang [19] and MGIT [9] benchmarks. For OTB99_Lang [19], inaccurate natural language descriptions in official annotations create interference for tracking, while concise text provides further assurance for BBox that already expresses temporal and spatial relationships well, reducing interference. For MGIT [9], the algorithm is unable to understand complex temporal relationships and can only extract semantic information from concise text. Official text annotations of LaSOT [3] lie between the two and are most conducive to the current algorithm, resulting in the best performance.

(2) **For short-term tracking task, dense concise text will bring greater gains. While dense detailed text is more suitable for the other two tasks.** Looking at the results of testing after retraining with different texts, dense concise text has the greatest impact on OTB99_Lang [19]. We think this is because the text provides precise object descriptions, further compensating for the shortcomings of BBox. The algorithm can further improve its performance on MGIT [9] by learning from dense detailed text, because they can provide high-level semantic information that BBox cannot exhibit, such as temporal and spatial relationships. By text updating that best suits the memory system of algorithm, we provide the algorithm with precise and timely high-level semantic information, which is more helpful for understanding long video.

(3) **The text processing method and multi-modal alignment ability need to be adjusted and improved.** The current algorithm cannot fully understand and learn complex temporal and spatial relationships. When the text processing and multi-modal alignment abilities of algorithm are adjusted and improved, text with more information will show even greater potential.

4.3. Visualization

As shown in Fig. 5, we visualize the tracking results of the retrained model with official and dense concise text on three challenging sequences from OTB99_Lang [19]. In these sequences, the official text annotations can only cover a short time for the changes in the object. The scenes contain distractors, and the appearance of the object undergoes significant changes. The retrained model exhibits greater robustness with dense concise text compared to the official one. This validates that our generated text helps tracker to address these challenges.

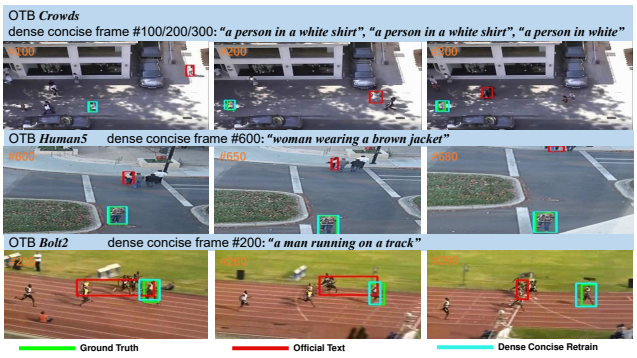

Figure 5. Visualization of tracking results on dense concise text annotations retrained algorithm.

5. Conclusions

Object tracking is the basis for advanced tasks such as video understanding, and VLT may offer a potential path for enhancing tracking capabilities. In this paper, we propose DTLLM-VLT, a unified prompt framework, and generate diverse multi-granularity text descriptions. We analyze the results under different natural language descriptions for three representative benchmarks, aiming to provide new insights for the evaluation of different tracking tasks.

In our perspective, enhancing algorithm performance requires a comprehensive understanding of the properties of the datasets. We explore how leveraging the generative capabilities of LLM can help us improve VLT datasets and provide a new analytical approach from a multi-modal perspective for the field of video understanding. We hope this work can be expanded to incorporate more datasets, thereby enhancing support for vision datasets understanding research.

6. Acknowledgement

This work is jointly supported by the National Science and Technology Major Project (No.2022ZD0116403); the National Natural Science Foundation of China (No.62176255); the Strategic Priority Research Program of Chinese Academy of Sciences (No.XDA27000000).

References

[1] Richard C Atkinson and Richard M Shiffrin. Human memory: A proposed system and its control processes. In *Psychology of learning and motivation*, pages 89–195. Elsevier, 1968. 5

[2] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023. 5

[3] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5369–5378, 2019. 1, 2, 3, 4, 6, 7, 8

[4] Heng Fan, Hexin Bai, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Harshit, Mingzhen Huang, Juehuan Liu, et al. Lasot: A high-quality large-scale single object tracking benchmark. *International Journal of Computer Vision*, 129: 439–461, 2021. 2, 3

[5] Qi Feng, Vitaly Ablavsky, Qinxun Bai, and Stan Sclaroff. Robust visual object tracking with natural language region proposal network. *arXiv preprint arXiv:1912.02048*, 1(7):8, 2019. 3

[6] Qi Feng, Vitaly Ablavsky, Qinxun Bai, Guorong Li, and Stan Sclaroff. Real-time visual object tracking with natural language description. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 700–709, 2020. 3

[7] Qi Feng, Vitaly Ablavsky, Qinxun Bai, and Stan Sclaroff. Siamese natural language tracker: Tracking by natural language descriptions with siamese trackers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5847–5856, 2021. 3

[8] Mingzhe Guo, Zhipeng Zhang, Heng Fan, and Liping Jing. Divert more attention to vision-language tracking. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 4446–4460, 2022. 3

[9] Shiyu Hu, Dailing Zhang, Meiqi Wu, Xiaokun Feng, Xuchen Li, Xin Zhao, and Kaiqi Huang. A multi-modal global instance tracking benchmark (mgit): Better locating target in complex spatio-temporal and causal relationship. In *Advances in Neural Information Processing Systems*, pages 25007–25030, 2023. 1, 2, 3, 4, 6, 7, 8

[10] Shiyu Hu, Xin Zhao, Lianghua Huang, and Kaiqi Huang. Global instance tracking: Locating target more like humans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):576–592, 2023. 3

[11] Shiyu Hu, Xin Zhao, and Kaiqi Huang. Sotverse: A user-defined task space of single object tracking. *International Journal of Computer Vision*, 132:872–930, 2024. 3

[12] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1562–1577, 2021. 3

[13] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 5

[14] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Luka Cehovin, Gustavo Fernandez, Tomas Vojir, Gustav Hager, Georg Nebehay, and Roman Pflugfelder. The visual object tracking vot2015 challenge results. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 1–23, 2015. 3

[15] Matej Kristan, Jiri Matas, Aleš Leonardis, Tomáš Vojíř, Roman Pflugfelder, Gustavo Fernandez, Georg Nebehay, Fatih Porikli, and Luka Čehovin. A novel performance evaluation methodology for single-target trackers. *IEEE transactions on pattern analysis and machine intelligence*, 38(11):2137–2155, 2016. 3

[16] Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, Luka ˇCehovin Zajc, Tomas Vojir, Goutam Bhat, Alan Lukezic, Abdelrahman Eldesokey, et al. The sixth visual object tracking vot2018 challenge results. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 3

[17] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kamarainen, Luka ˇCehovin Zajc, Ondrej Drbohlav, Alan Lukezic, Amanda Berg, et al. The seventh visual object tracking vot2019 challenge results. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 3

[18] Yihao Li, Jun Yu, Zhongpeng Cai, and Yuwen Pan. Cross-modal target retrieval for tracking by natural language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4931–4940, 2022. 3

[19] Zhenyang Li, Ran Tao, Efstratios Gavves, Cees GM Snoek, and Arnold WM Smeulders. Tracking by natural language specification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6495–6503, 2017. 1, 2, 3, 4, 5, 6, 7, 8

[20] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European conference on computer vision (ECCV)*, pages 300–317, 2018. 3

[21] Gabriel A Radvansky. *Human memory*. Routledge, 2021. 5

[22] Rael D Strous, Nelson Cowan, Walter Ritter, and Daniel C Javitt. Auditory sensory ("echoic") memory dysfunction in schizophrenia. *The American journal of psychiatry*, 152(10): 1517–1519, 1995. 5

[23] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 5

[24] Jack Valmadre, Luca Bertinetto, Joao F Henriques, Ran Tao, Andrea Vedaldi, Arnold WM Smeulders, Philip HS Torr, and Efstratios Gavves. Long-term tracking in the wild: A bench-

mark. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–685, 2018. 3

[25] Rong Wang, Zongheng Tang, Qianli Zhou, Xiaoqian Liu, Tianrui Hui, Quange Tan, and Si Liu. Unified transformer with isomorphic branches for natural language tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 3

[26] Xiao Wang, Chenglong Li, Rui Yang, Tianzhu Zhang, Jin Tang, and Bin Luo. Describe and attend to track: Learning natural language guided structural representation and visual attention for object tracking. *arXiv preprint arXiv:1811.10014*, 2018. 3

[27] Xiao Wang, Xiujun Shu, Zhipeng Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng Wu. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13763–13773, 2021. 2, 3, 4, 5

[28] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2411–2418, 2013. 2

[29] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(09):1834–1848, 2015. 2, 3

[30] Zhengyuan Yang, Tushar Kumar, Tianlang Chen, Jingsong Su, and Jiebo Luo. Grounding-tracking-integration. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(9):3433–3443, 2021. 3

[31] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. *arXiv preprint arXiv:2312.10032*, 2023. 5

[32] Haojie Zhao, Xiao Wang, Dong Wang, Huchuan Lu, and Xiang Ruan. Transformer vision-language tracking via proxy token guided cross-modal fusion. *Pattern Recognition Letters*, 168:10–16, 2023. 3

[33] Xin Zhao, Shiyu Hu, Yipei Wang, Jing Zhang, Yimin Hu, Rongshuai Liu, Haibin Ling, Yin Li, Renshu Li, Kun Liu, et al. Biodrone: A bionic drone-based single object tracking benchmark for robust vision. *International Journal of Computer Vision*, pages 1–26, 2023. 3

[34] Yaozong Zheng, Bineng Zhong, Qihua Liang, Guorong Li, Rongrong Ji, and Xianxian Li. Towards unified token learning for vision-language tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 2, 3, 6, 7

[35] Li Zhou, Zikun Zhou, Kaige Mao, and Zhenyu He. Joint visual grounding and tracking with natural language specification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23151–23160, 2023. 3