Visual Language Tracking with Multi-modal Interaction: A Robust Benchmark

Xuchen Li[1,2]   **Shiyu Hu**[3]   **Xiaokun Feng**[1,2]   **Dailing Zhang**[1,2]
Meiqi Wu[4]   **Jing Zhang**[1]   **Kaiqi Huang**[1,2,5]

[1]CRISE, Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]School of Physical and Mathematical Sciences, Nanyang Technological University
[4]School of Computer Science and Technology, University of Chinese Academy of Sciences
[5]CAS Center for Excellence in Brain Science and Intelligence Technology
lixuchen2024@ia.ac.cn, shiyu.hu@ntu.edu.sg {fengxiaokun2022, zhangdailing2023}@ia.ac.cn,
wumeiqi18@mails.ucas.ac.cn, {jing_zhang, kqhuang}@ia.ac.cn

Abstract

Visual Language Tracking (VLT) enhances tracking by mitigating the limitations of relying solely on the visual modality, utilizing high-level semantic information through language. This integration of the language enables more advanced human-machine interaction. The essence of interaction is cognitive alignment, which typically requires multiple information exchanges, especially in the sequential decision-making process of VLT. However, current VLT benchmarks do not account for multi-round interactions during tracking. They provide only an initial text and bounding box (bbox) in the first frame, with no further interaction as tracking progresses, deviating from the original motivation of the VLT task. To address these limitations, we propose a novel and robust benchmark, **VLT-MI** (**V**isual **L**anguage **T**racking with **M**ulti-modal **I**nteraction), which introduces multi-round interaction into the VLT task for the first time. (1) We generate diverse, multi-granularity texts for multi-round, multi-modal interaction based on existing mainstream VLT benchmarks using DTLLM-VLT, leveraging the world knowledge of LLMs. (2) We propose a new VLT interaction paradigm that achieves multi-round interaction through text updates and object recovery. When multiple tracking failures occur, we provide the tracker with more aligned texts and corrected bboxes through interaction, thereby expanding the scope of VLT downstream tasks. (3) We conduct comparative experiments on both traditional VLT benchmarks and VLT-MI, evaluating and analyzing the accuracy and robustness of trackers under the interactive paradigm. This work offers new insights and paradigms for the VLT task, enabling a fine-grained evaluation of multi-modal trackers. We believe this approach can be extended to additional datasets in the future, supporting broader evaluations and comparisons of video-language model capabilities.

1 Introduction

Single Object Tracking (SOT) has progressed from short-term tracking [1–3] to long-term tracking [4], and more recently to global instance tracking [5]. Researchers are increasingly focusing on making trackers more human-like through interaction and assessing tracking intelligence based on human visual capabilities [5, 6]. However, due to the limitations of single-modality systems, SOT faces challenges in intuitively interacting with humans, which impedes its progression toward more complex downstream tasks. Visual Language Tracking (VLT) [7, 8] introduces the textual modality, aiming to facilitate more flexible and intuitive human interactions, thus enabling the evaluation of

Table 1: Comparison of current datasets for VLT. VLT-MI is the first VLT benchmark that supports multi-round multi-modal interaction, covering three different tracking tasks. "STT", "LTT" and "GIT" refer to Short-term Tracking, Long-term Tracking and Global Instance Tracking.

Dataset	Video number	Min frame	Mean frame	Max frame	Total frames	Tracking task	Multi-round Multi-modal Interaction
OTB99_Lang	99	71	590	3,872	59K	STT	×
LaSOT	1,400	1,000	2,506	11,397	3.52M	LTT	×
TNL2K	2,000	21	622	18,488	1.24M	STT	×
MGIT	150	4,008	14,920	29,834	2.03M	GIT	×
VLT-MI	3,619	21	1,824	29,834	6.60M	STT & LTT & GIT	✓

human-like intelligence in trackers. Despite this goal, current VLT trackers [9–15] do not incorporate multi-round, multi-modal interactions during the tracking process. Both training and testing typically provide only a text description and bounding box in the initial frame, which deviates from the original motivation of the VLT task. Moreover, most existing VLT benchmarks [7, 8, 16] annotate texts at a single granularity, focusing mainly on object features in the initial frame. This approach hampers the ability to provide high-quality information through multi-round interactions, leading to misalignment between modalities. At the evaluation level, current benchmark metrics broadly reflect tracker performance but lack fine-grained assessments of robustness.

To address these issues, we develop the VLT-MI benchmark, which allows for dynamic interaction during the tracking process, enabling text updates and object recovery based on the evolving context of the video. We utilize DTLLM-VLT [17] to generate high-quality video textual information for interaction, facilitating improved alignment between visual and textual modalities throughout the tracking process. Additionally, we introduce more detailed, interaction-specific metrics to provide a deeper understanding of how effectively the tracker maintains object continuity and adapts to situational changes, thus offering a more nuanced evaluation of tracker robustness and intelligence.

2 Construction of VLT-MI

We utilize DTLLM-VLT [17] to provide concise and detailed textual descriptions for videos in four mainstream VLT datasets every 100 frames. As shown in Table 1, VLT-MI is the first VLT benchmark with multi-round multi-modal interactions. During the interaction process, if the bbox Intersection over Union (IoU) between the predicted object and the ground truth across 10 consecutive frames falls below 0.5, we consider the tracker to have lost accurate tracking of the object. In such cases, interaction is required to guide the tracker. This interaction is achieved by updating the textual information closest to the interaction frame and providing an accurate bounding box for the current frame. To enhance the quality of textual information during interaction, we generate both concise and detailed descriptions. The concise text focuses solely on the essential features of the object, while the detailed text also incorporates background descriptions. The interaction process is illustrated in Figure 1.

Figure 1: Example of tracking with multi-round multi-modal interaction. We provide the tracker with the correct bbox and a more accurately aligned concise/detailed text through interaction.

3 Experimental Results

We selected four representative datasets—OTB99_Lang [7], LaSOT [4], TNL2K [8], and MGIT [16]—to evaluate short-term, long-term, and global instance tracking tasks. For our baseline model, we chose the state-of-the-art VLT tracker, JointNLT [10], and evaluated its performance across all four benchmarks. To ensure a fair comparison of tracking performance, we directly tested the model using the official weights, assessing precision (PRE), normalized precision (N-PRE), and success rate (SR) [5]. Additionally, we reported the average multi-modal interaction number (AMI), relative average multi-modal interaction number (R-AMI), average maximum tracking success length (AMSL), and relative average maximum tracking success length (R-AMSL) to measure the tracker's robustness. The results are presented in Table 2, Table 3, and Figure 2.

Table 2: Comparison of tracking accuracy on VLT-MI. The best two results are highlighted in red and blue.

JointNLT [10]	OTB99_Lang [7]			MGIT [16]			LaSOT [4]			TNL2K [8]		
	PRE	N-PRE	SR	PRE	N-PRE	SR	PRE	N-PRE	SR	PRE	N-PRE	SR
Traditional	65.0	54.1	50.5	29.8	66.4	47.7	54.0	55.2	53.2	48.7	52.1	50.0
Interaction_Concise	64.8	53.1	49.3	29.2	63.1	48.3	56.2	58.2	56.3	48.1	52.0	49.9
Interaction_Detailed	63.7	52.3	48.8	29.0	62.9	48.0	55.1	57.8	56.1	48.2	52.3	50.2

Table 3: Comparison of robustness on VLT-MI. We report AMI and AMSL.

JointNLT [10]	OTB99_Lang [7]		MGIT [16]	
	AMI	AMSL	AMI	AMSL
Interaction_Concise	8	118	229	2586
Interaction_Detailed	10	150	235	2661
JointNLT [10]	LaSOT [4]		TNL2K [8]	
	AMI	AMSL	AMI	AMSL
Interaction_Concise	37	431	13	171
Interaction_Detailed	40	455	13	163

Tracking Accuracy. The original intention of interacting with the tracker is to restore tracking when the tracker fails over an extended period, thereby improving tracking accuracy. However, while tracking accuracy on LaSOT [4] meets expectations, performance on several other benchmarks has declined after the interaction. We believe this is because the tracker is not yet fully adapted to the multi-round interaction mode and is overly sensitive to the textual input. This behavior is similar to completing the VLT task by "memorizing the answers."

Robustness. We evaluate the robustness of the interactive tracking algorithm from two perspectives: the average number of interactions and the average maximum length of successful tracking sequences. As shown in Table 3 and Figure 2, the number of interactions gradually increases as task difficulty rises and sequence length extends. The number of interactions ranges from 8 to 13 for short-term tracking, to 37 for long-term tracking, and up to 229 for global instance tracking. R-AMI reflects the proportion of interaction frames within the sequence, while R-AMSL represents the percentage of the longest successful tracking subsequence in the sequence. As tracking progresses from short-term to long-term and global instance tracking, the proportion of interaction frames steadily decreases. Among the four benchmarks, MGIT [16] exhibits the highest proportion of longest successful tracking subsequences, whereas OTB99_Lang [7] has the lowest, which is consistent with expectations.

R-AMI (%)	OTB99_Lang	MGIT	LaSOT	TNL2k
Detailed	1.7	1.18	1.67	1.69
Concise	1.52	1.28	1.62	1.73

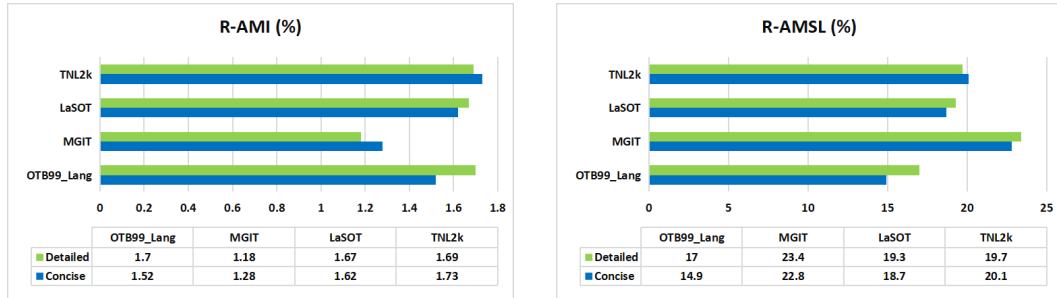| R-AMSL (%)	OTB99_Lang	MGIT	LaSOT	TNL2k
Detailed	17	23.4	19.3	19.7
Concise	14.9	22.8	18.7	20.1

Figure 2: Comparison of robustness on VLT-MI with R-AMI and R-AMSL. The calculation of relative metrics is based on absolute metrics, divided by the sequence length and then averaged.

4 Conclusions

VLT extends the SOT task by introducing a textual modality, which naturally enhances the interactive capabilities between the tracker and humans. In this paper, we present VLT-MI, the first implementation of multi-round, multi-modal interaction within object tracking. Interactions are facilitated through textual updates and target recovery when the tracker repeatedly fails to follow a specific object. We analyze interactive behaviors and robustness, aiming to provide new insights into the advancement of visual language trackers.

From our perspective, human-computer interaction is a critical objective for video language tasks, as demonstrated by VLT. We explore how to integrate human factors into video language tasks to support multi-modal interaction, and we introduce a novel robustness evaluation method from an interaction standpoint. We hope that this work can be extended to more video language tasks, further advancing the development of video language models.

References

[1] 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] 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] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] Yinchao Ma, Yuyang Tang, Wenfei Yang, Tianzhu Zhang, Jinpeng Zhang, and Mengxue Kang. Unifying visual and vision-language tracking via contrastive learning. *arXiv preprint arXiv:2401.11228*, 2024.

[12] 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.

[13] 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.

[14] Mingzhe Guo, Zhipeng Zhang, Heng Fan, and Liping Jing. Divert more attention to vision-language tracking. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Proceedings of the Advances in Neural Information Processing Systems*, volume 35, pages 4446–4460, 2022.

[15] 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.

[16] 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*, volume 36, pages 25007–25030, 2023.

[17] Xuchen Li, Xiaokun Feng, Shiyu Hu, Meiqi Wu, Dailing Zhang, Jing Zhang, and Kaiqi Huang. Dtllm-vlt: Diverse text generation for visual language tracking based on llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7283–7292, 2024.