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

CVPR SEATTLE, WA JUNE 17-21, 2024

Platform for more information

中国科学院自动化研究所 INSTITUTE OF AUTOMATION CHINESE ACADEMY OF SCIENCES
中国科学院大学 University of Chinese Academy of Sciences
RISE 智能系统与工程研究中心 Center for Research on Intelligent System and Engineering

Motivation

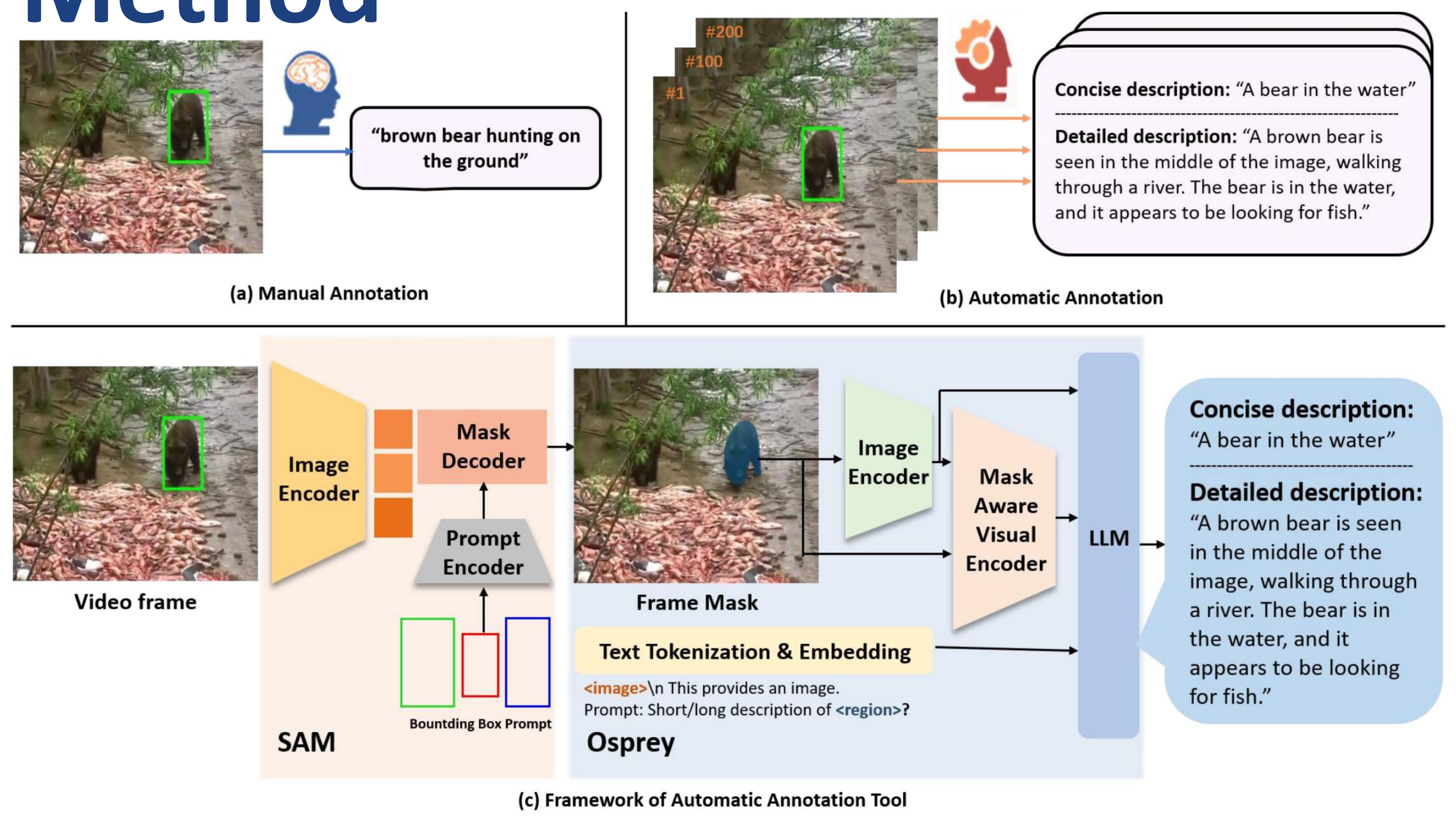
Most VLT benchmarks are annotated in a single granularity and lack a coherent semantic framework to provide scientific guidance.

Video: Annotations mainly describe the first frame, which may misguide the algorithm.

Language: Environment is complex and variable. Different VLT datasets lack a coherent framework.

Comparison of different text annotations, video length, and content on three benchmarks, most of VLT benchmark suffer from issues of inconsistent text styles and single annotation granularity.

Method



Video — Language

1.9M words
14.8K non-repetitive words.
7,238 initial descriptions
128.4K dense descriptions

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 video understanding.

Laborious and time-consuming human annotators for high-quality annotations. → Method based on LLM, efficiently generating high-quality scientific text at scale. It can seamlessly apply to various tasks.

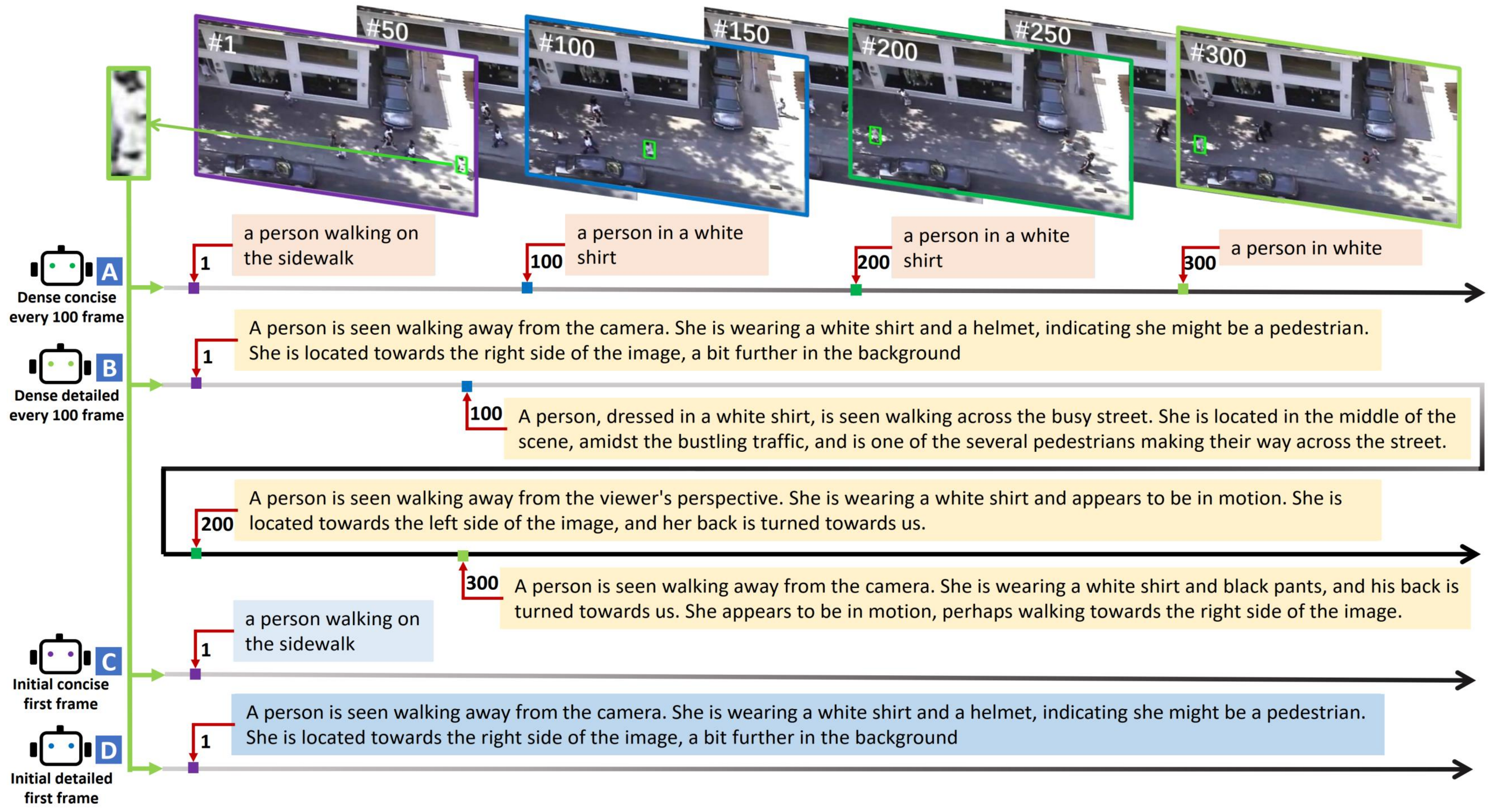
Contributions

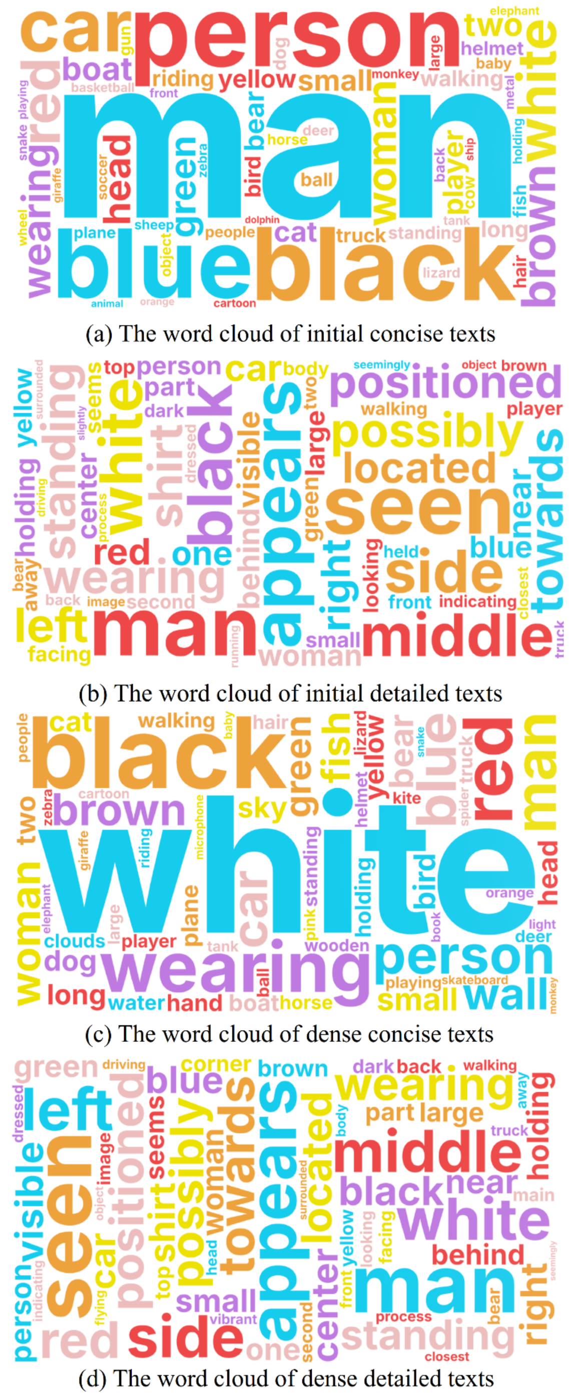
- We develop DTLLM-VLT, a model based on LLM, aimed at efficiently generating high-quality scientific text for tracking datasets at scale. DTLLM-VLT can seamlessly apply to various tracking tasks.
- We generate diverse text for three prominent VLT benchmarks, addressing four levels of granularity. This 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.

Diverse Text Generation

Multi-Granularity Diverse Semantic Generation Strategy

Four Granularity Evaluation Mechanism





(a) The word cloud of initial concise texts
(b) The word cloud of initial detailed texts
(c) The word cloud of dense concise texts
(d) The word cloud of dense detailed texts

Experiments

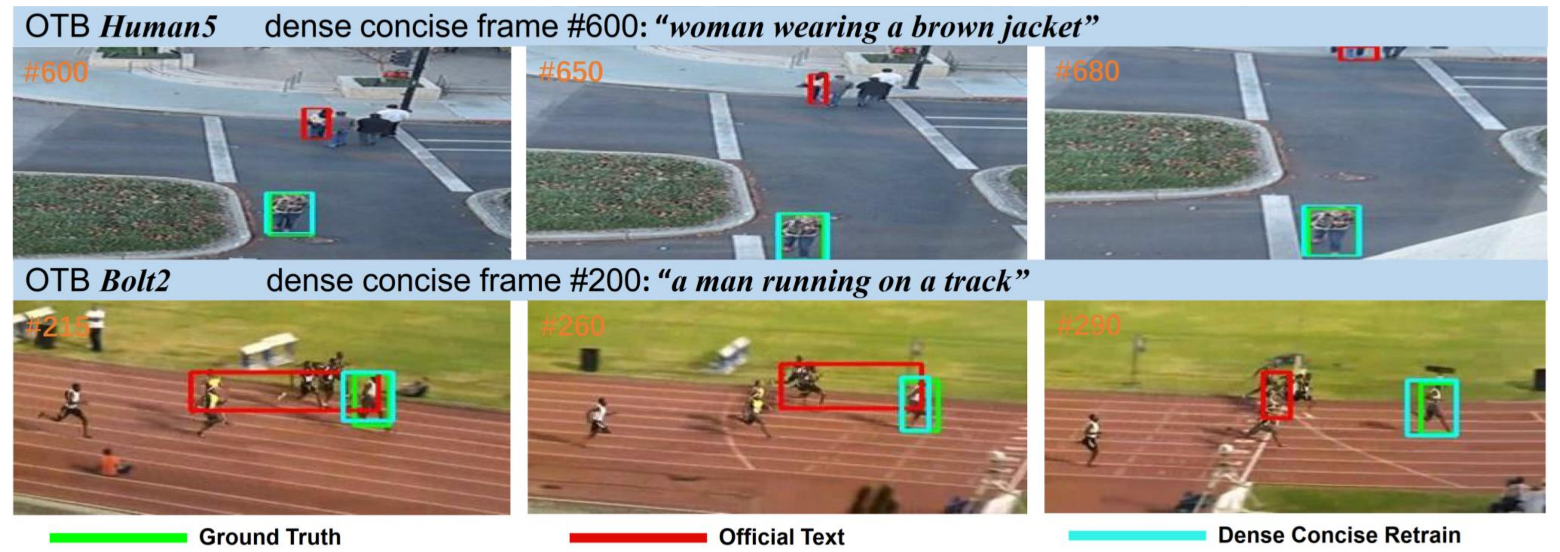
Comparison with testing directly

Method	OTB99_Lang			MGIT			LaSOT		
	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

Comparison with retraining and testing

Method	OTB99_Lang			MGIT			LaSOT		
	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

Visualization



Conclusion

- The existing algorithm tends to learn and understand short text.
- For short-term tracking task, dense concise text will bring greater gains. While dense detailed text is more suitable for the other two tasks.
- The text processing method and multi-modal alignment ability need to be adjusted and improved.