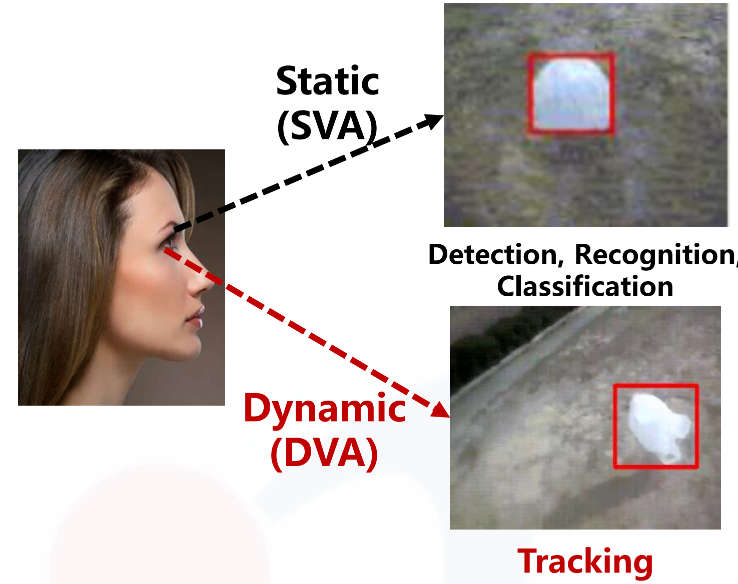
DTLLM-VLT: Diverse Text Generation for Visual Language Tracking Based on LLM

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

Task Definition

Humans are "visual creatures"



Object tracking is a fundamental function of the human dynamic visual system

modeling

VLT (Visual Language Tracking)

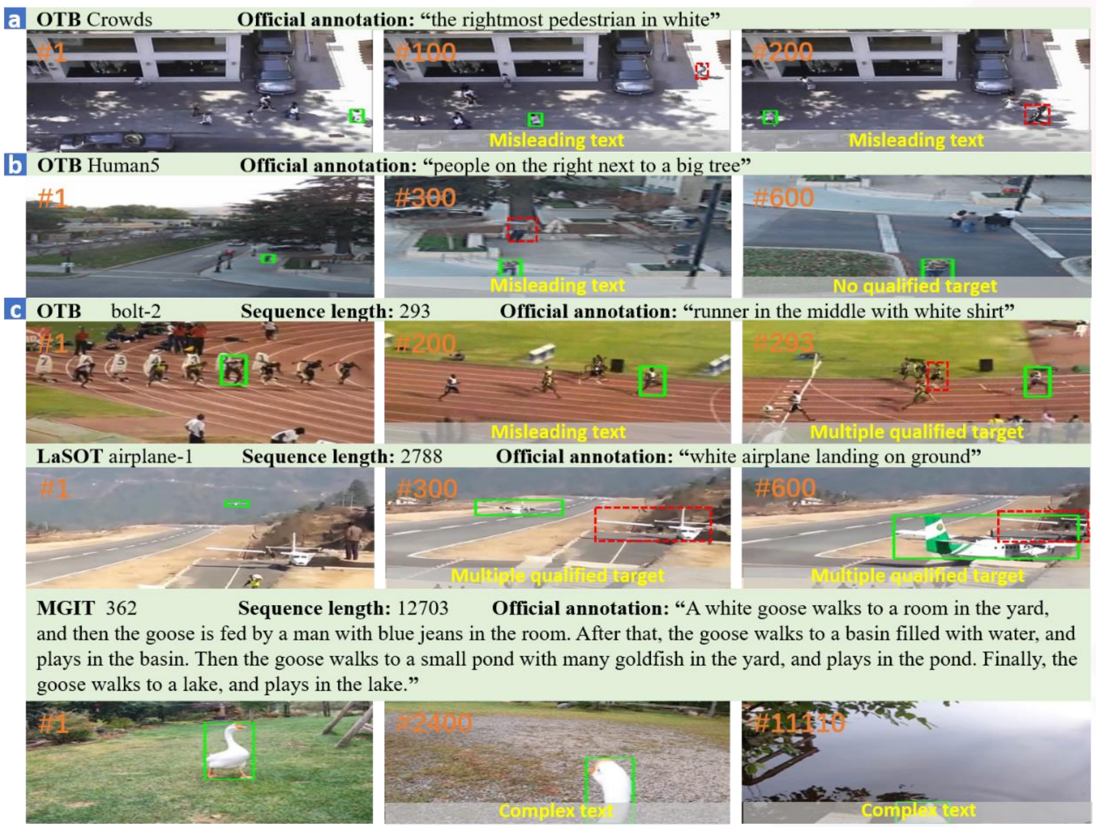
- **Definition :** Only providing the initial position and natural language descriptions of a moving object and continuously locating it in a video sequence.

First frame : Initialize by NL & BBox

Video sequence : Locate continuously in video sequence

Motivation

- Most VLT benchmarks are annotated in a **single granularity** and **lack a coherent semantic framework** to provide scientific guidance.
- Current VLT benchmarks considers studying from different perspective :
  - **Limitations1.** Semantic annotations in OTB99_Lang mainly describe **the first frame**, which may **misguide the algorithm**.
  - **Limitations2.** Sequence in MGIT has such **complex text** that they are **not conducive to algorithmic learning**.

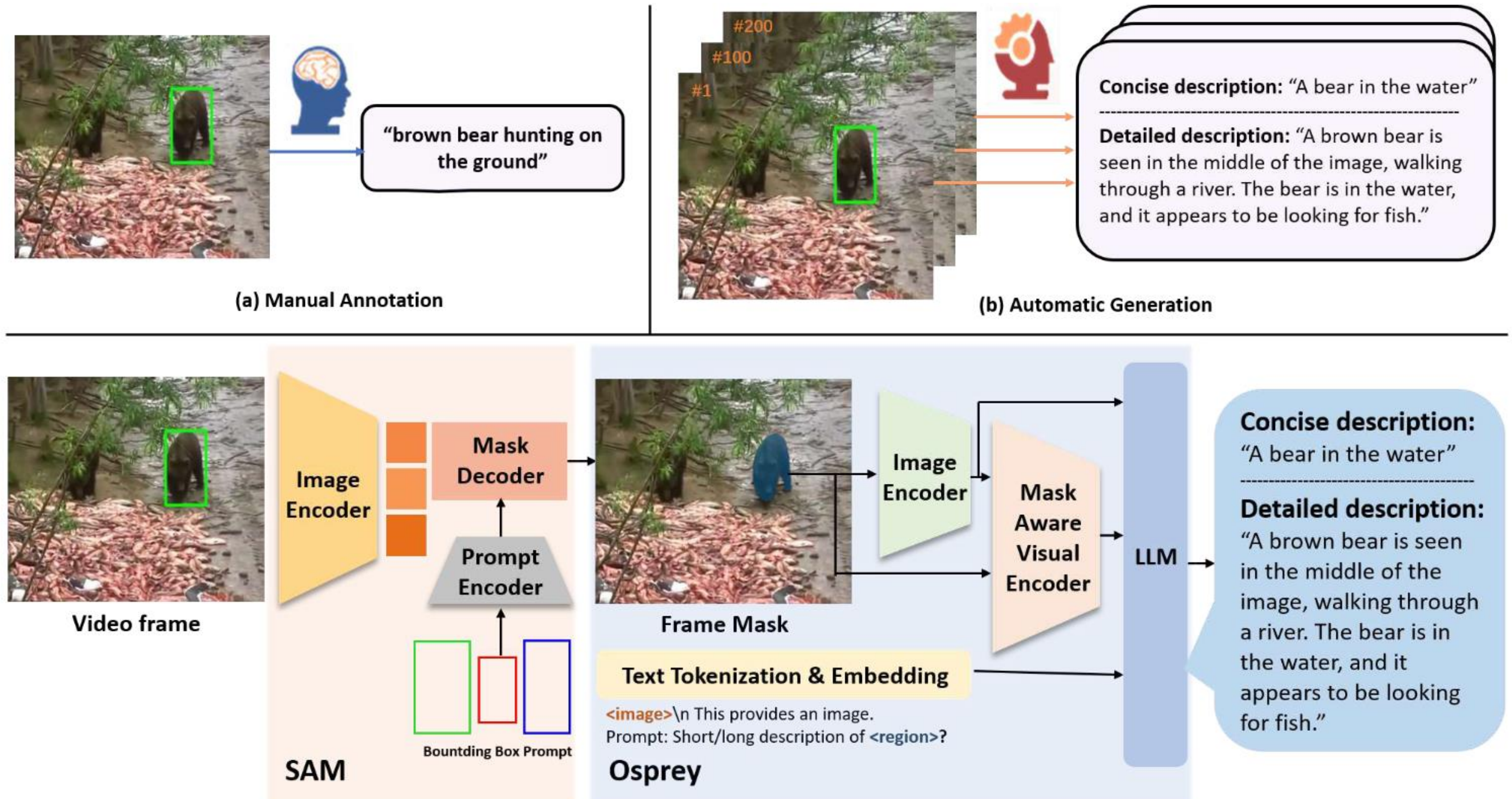


- **Research objective :** Using LLM to provide multi-granularity semantic information for VLT from efficient and diverse perspectives, enabling fine-grained evaluation. This work can be extended to more datasets to support vision datasets understanding.

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

Contribution 1: Diverse text generation method based on LLM (DTLLM-VLT)

- **Diverse texts** matter → Integrating the **LLM** into the text generation process, offer a **diverse environment** conducive to VLT research.

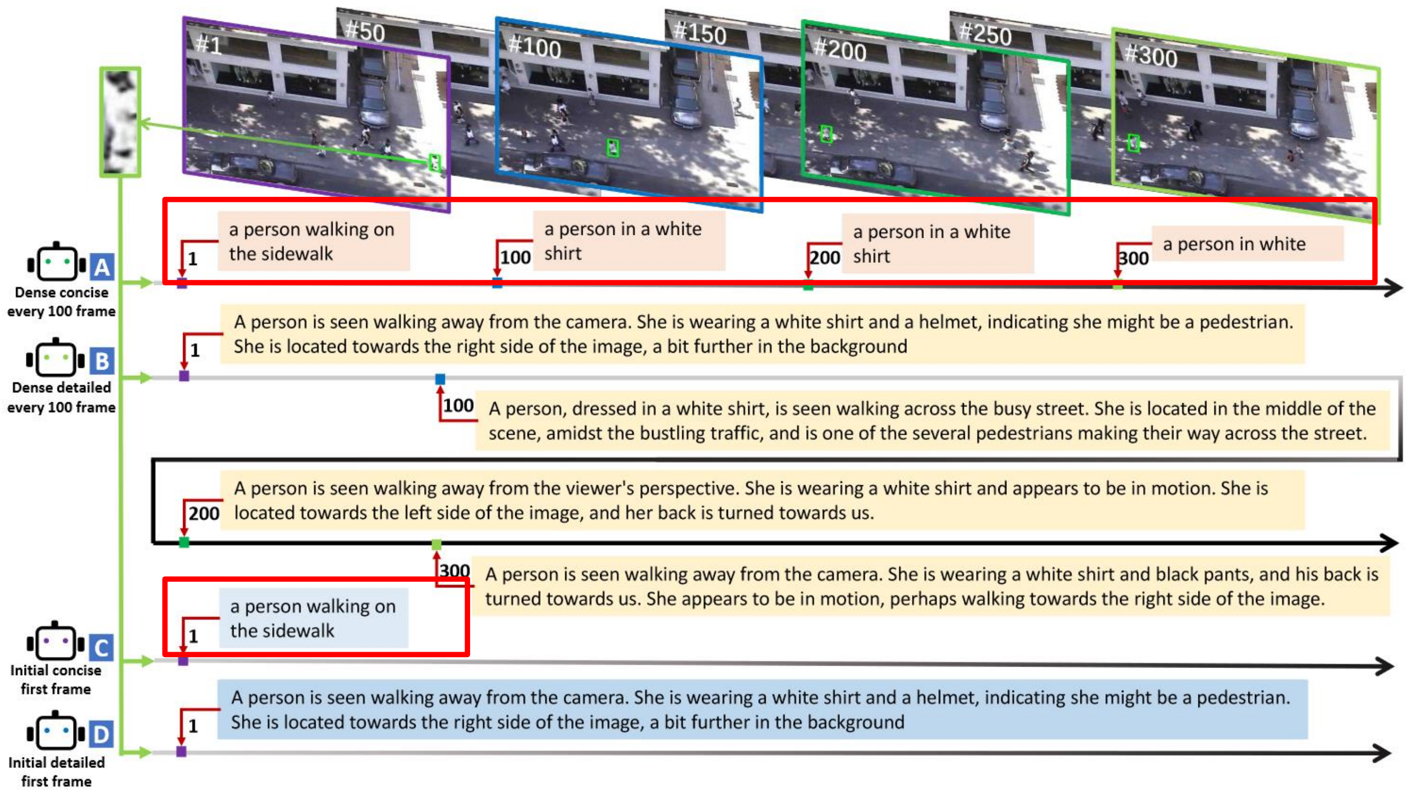


(c) Framework of DTLLM-VLT

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

Contribution 2: Multi-Granularity Diverse Semantic Generation Strategy

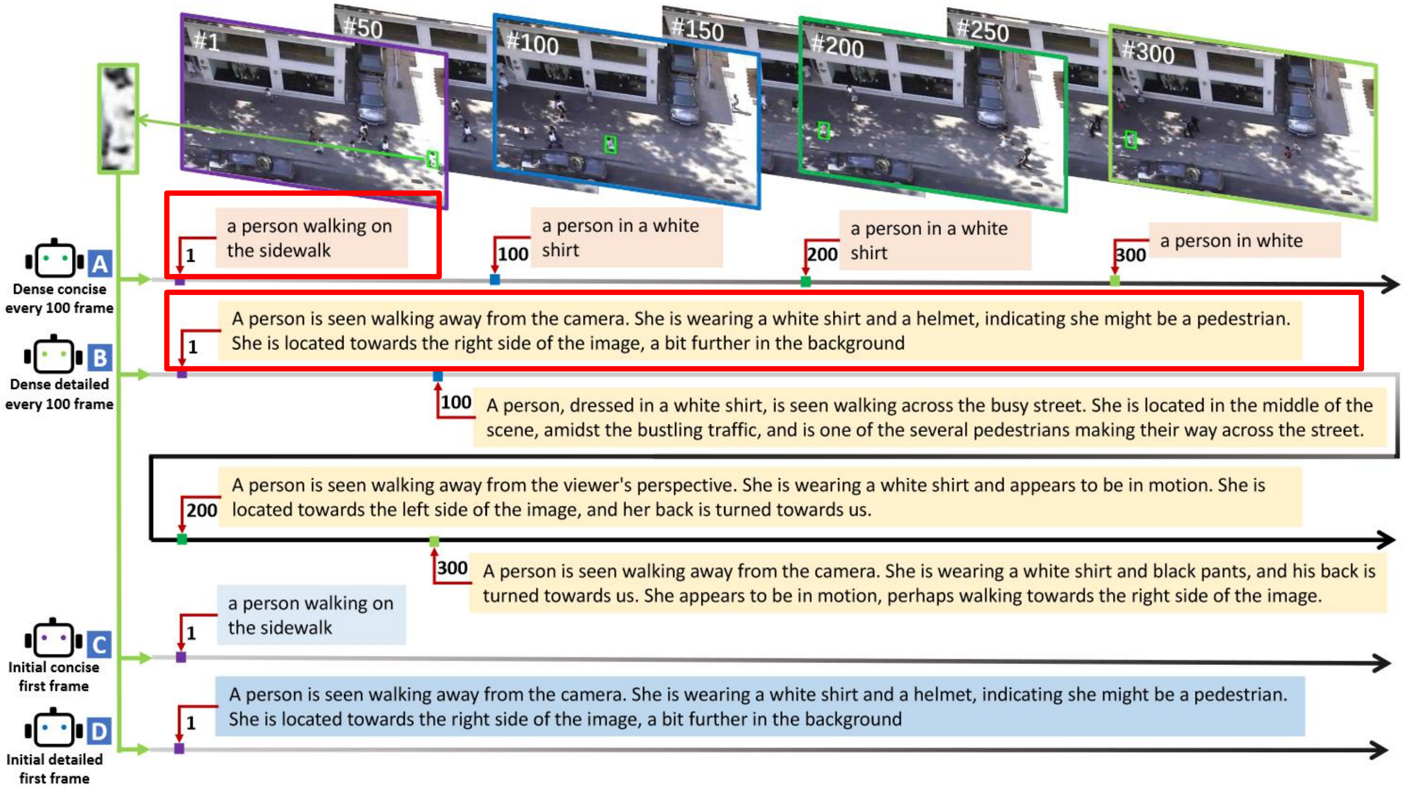
- Applying **multi-granularity generation**
  - **Initial texts:** Following the text annotations method in OTB99_Lang and TNL2K, we generate text for the **initial frame** of each video.
  - **Dense texts**: Considering the worst situation and infer that the algorithm lacks an efficient memory system. Consequently, at 25 FPS, equating to **every 100 frames** in 4 seconds, we supply the algorithm with relevant generated text.



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

Contribution 2: Multi-Granularity Diverse Semantic Generation Strategy

- Applying **multi-granularity generation**
  - **Concise texts:** 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.
  - **Detailed texts**: 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.



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

Contribution 2: Multi-Granularity Diverse Semantic Generation Strategy

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

Dataset	Number of Language Description				
	Official	*Dense Concise*	*Dense Detailed*	*Initial Concise*	*Initial Detailed*
OTB99_Lang	99	*596*	*596*	*99*	*99*
LaSOT	1,400	*35.2K*	*35.2K*	*1,400*	*1,400*
TNL2K	2,000	*12.4K*	*12.4K*	*2,000*	*2,000*
MGIT	1,753	*16.1K*	*16.1K*	*120*	*120*

(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

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

Contribution 3: Evaluation mechanism and analysis for VLT task

- Mechanism A: Utilizing the official weight files provided, we keep all parameters unchanged and **directly test** the tracking performance.

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

- Mechanism B: We **continue training** for an additional 50 epochs based on the official weights, using datasets such as OTB99_Lang, LaSOT, TNL2K, and RefCOCOg. During the training process, we replace the official texts with different texts.

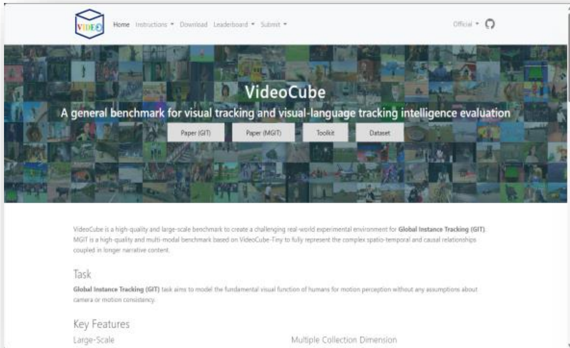
Method	OTB99_Lang [19]			MGIT [9]			LaSOT [3]		
	AUC	P_{Norm}	P	AUC	P_{Norm}	P	AUC	P_{Norm}	P
Official	69.0	82.0	89.5	73.5	77.2	54.3	69.9	82.2	75.7
Initial Concise	70.0	84.3	90.5	73.6	77.4	54.2	69.6	81.8	75.4
Initial Detailed	70.3	85.6	91.4	74.1	78.3	54.5	69.4	81.5	75.1
Dense Concise	71.3	86.0	92.5	74.0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

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

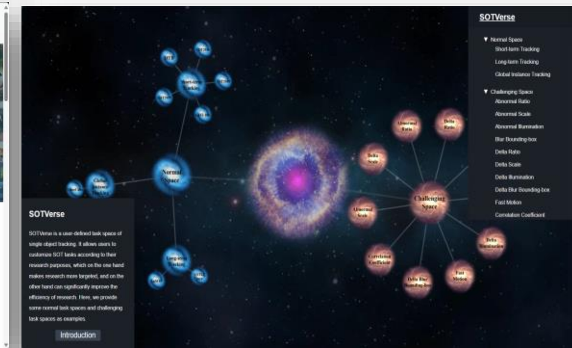
Experimental analysis: Diverse texts are suitable for different tasks

- 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.

Series Work on SOT and VLT

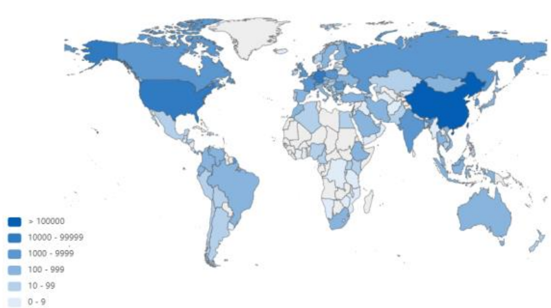


VideoCube Platform



SOTVerse Platform

GOT-10k Platform



3.9M IP Views from 160+ Countries

7,000+ Downloads and Evaluation for 215k+ Trackers

VideoCube Platform (TPAMI'23 & NIPS'23): http://videocube.aitestunion.com/
SOTVerse Platform (IJCV'23): http://metaverse.aitestunion.com/
GOT-10k Platform (TPAMI'21): http://got-10k.aitestunion.com/

Thanks!

CRISE @ CASIA