JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Joint Visual Grounding and Tracking with Natural Language Specification

Li Zhou[1], Zikun Zhou[2,1*], Kaige Mao[1], and Zhenyu He[1,*]
[1]Harbin Institute of Technology, Shenzhen [2]Peng Cheng Laboratory
lizhou.hit@gmail.com zhouzikunhit@gmail.com maokaige.hit@gmail.com zhenyuhe@hit.edu.cn

[Paper] [Project Page][Poster][Demo]

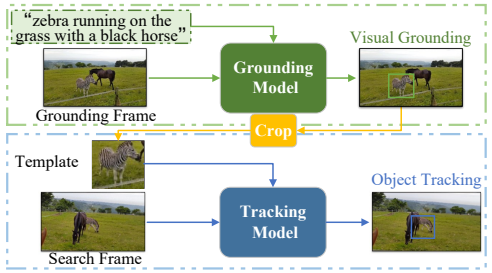
Speaker: Li Zhou

Overview

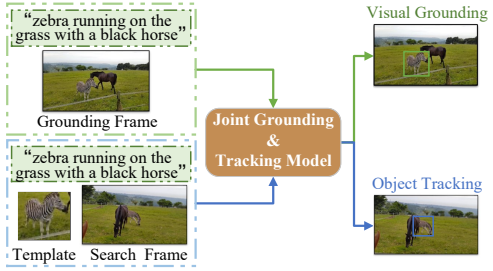
Motivation and Contribution

Motivation:

- Existing algorithms address this problem using a separate framework;
- Such a separated framework overlooks the link between visual grounding and tracking.



(a) Separated visual grounding and tracking framework



(b) Joint visual grounding and tracking framework

Key Contributions:

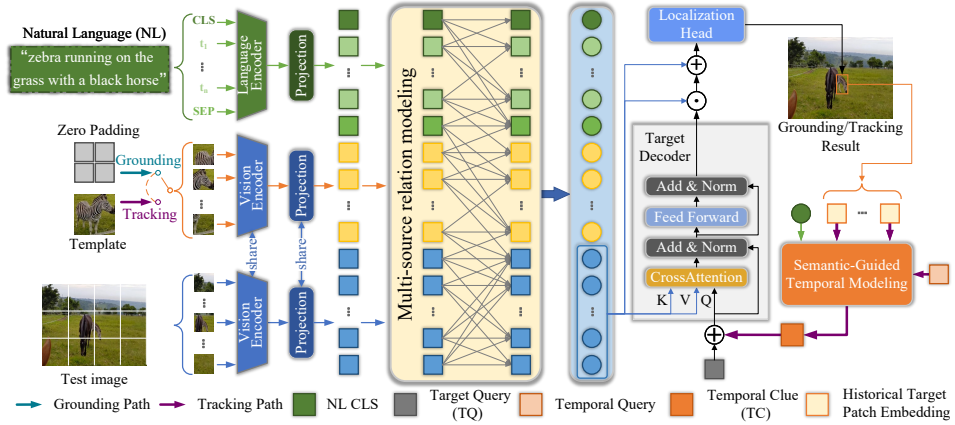
- Propose a joint visual grounding and tracking framework which accommodate the different references of the grounding and tracking processes;
- Propose a semantics-guided temporal modeling module to provide a temporal clue based on historical predictions for our joint model
- Achieve favorable performance against state-of-the-art algorithms on three natural language tracking datasets and one visual grounding dataset.

More details at: https://github.com/lizhou-cs/JointNLT

Proposed Model

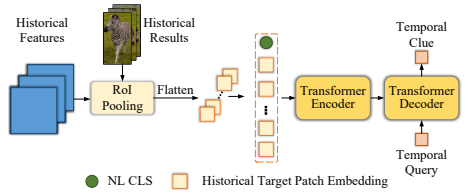
Overview of our joint visual grounding and tracking framework:

- Consists of modality encoder, multi-source relation modeling, target decoder, SGTM module and localization head.
- Integrates the processes of grounding and tracking by unifying the relation modeling of diverse references.



Architecture of the proposed SGTM:

- Utilize the language context and historical target status

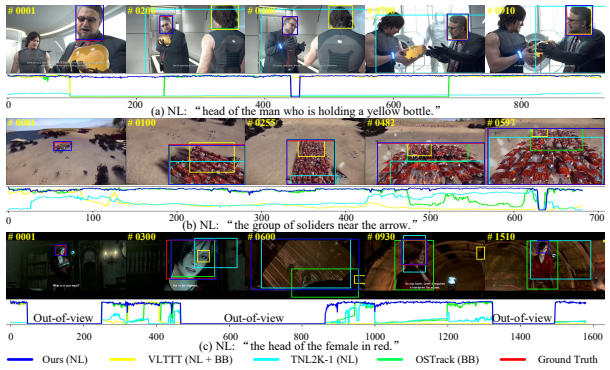


Pipeline:

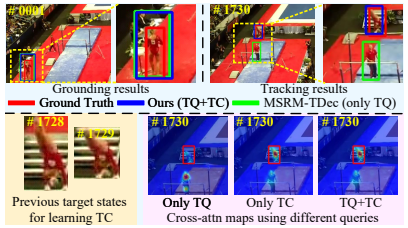
- Given a sequence and a natural language description, we first feed the description, the first frame (test image), and zero padding tokens into the model for visual grounding and accordingly obtain the template image.
- For each subsequent frame (test image), we feed it with the description and template image into the model for tracking.

Visualization Result

Qualitative comparison on three challenging sequences.



Visualization for revealing what temporal clue SGTM learns.



- We visualize the cross-attention maps in the target decoder and the prediction results
- With the help of TQ, model shows greater robustness to appearance variation

Results on Multi Benchmarks

State-of-the-art comparison on multi datasets

- The methods are separately divided into three categories based on initialization methods

Algorithms	Initialize	OTB99 AUC | PRE	LaSOT AUC | PRE	TNL2K AUC | PRE
AutoMatch	BB	–	0.583 | 0.599	0.472 | 0.435
TrDiMP	BB	–	0.639 | 0.663	0.523 | **0.528**
TransT	BB	–	0.649 | 0.690	0.507 | 0.517
STARK	BB	–	0.671 | 0.712	–
KeepTrack	BB	–	0.671 | 0.702	–
SwinTrack-B	BB	–	0.696 | 0.741	–
OSTrack-384	BB	–	**0.711** | **0.776**	**0.559** | –
TNLS-II	NL	0.250 | 0.290	–	–
RTTNLD	NL	0.540 | **0.780**	0.280 | 0.280	–
GTI	NL	0.581 | 0.732	0.478 | 0.476	–
TNL2K-1	NL	0.190 | 0.240	0.510 | 0.490	0.110 | 0.060
CTRNLT	NL	0.530 | 0.720	0.520 | 0.510	0.140 | 0.090
Ours	NL	**0.592** | 0.776	**0.569** | **0.593**	**0.546** | **0.550**
TNLS-III	NL+BB	0.550 | 0.720	–	–
RTTNLD	NL+BB	0.610 | 0.790	0.350 | 0.350	0.250 | 0.270
TNL2K-2	NL+BB	0.680 | 0.880	0.510 | 0.550	0.420 | 0.420
SNLT	NL+BB	0.666 | 0.804	0.540 | 0.576	0.276 | 0.419
VLTTT	NL+BB	**0.764** | **0.931**	**0.673** | **0.721**	0.531 | 0.533
Ours	NL+BB	0.653 | 0.856	0.604 | 0.636	**0.569** | **0.581**

Comparisons between separated and joint methods

		Separated Model VLTVG+STARK	Separated Model VLTVG+OSTrack	Separated Model SepRM	Joint Model Ours
FLOPs	Grounding	39.6G	39.6G	34.7G	34.9G
	Tracking	20.4G	48.3G	38.5G	42.0G
Time	Grounding	28.2ms	28.2ms	26.4ms	34.8ms
	Tracking	22.9ms	8.3ms	20.6ms	25.3ms
Params	Total	169.8M	214.7M	214.4M	153.0M
AUC	LaSOT	0.446	0.524	0.518	0.569
	TNL2K	0.373	0.399	0.491	0.546

Ablation experiment

Variants	LaSOT AUC	LaSOT PRE	TNL2K AUC	TNL2K PRE
SepRM	0.518	0.512	0.491	0.471
MSRM	0.536	0.550	0.511	0.500
MSRM-TDec	0.549	0.567	0.524	0.514
MSRM-TM	0.561	0.581	0.541	0.540
Our model	0.569	0.593	0.546	0.550

Comparison of our method with state-of-the-art algorithms for visual grounding

Algorithms	NMTree	LBYL-Net	ReSC-Large	TransVG	VLTVG	Ours
Accuracy	0.6178	0.6270	0.6312	0.6702	**0.7298**	0.7007

Conclusion

Conclusion: We propose a joint visual grounding and tracking framework by unifying the relation modeling. Besides, we propose a semantics-guided temporal modeling module modeling the historical target states with global semantic information as guidance.

Outline

Background & Motivation

➢ Task definition:

Tracking by BBox

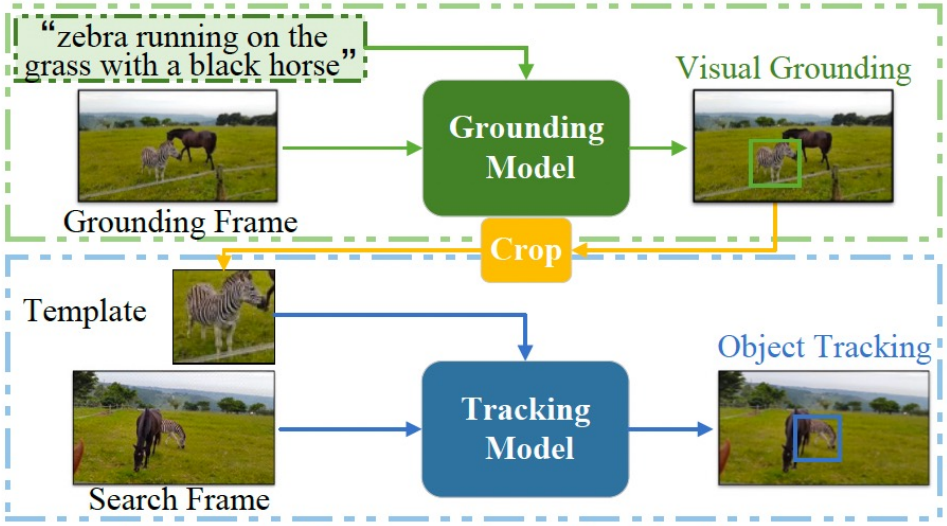
Tracking by Natural Language (NL)

- The bounding box only provides a static representation of the target state.
- The bounding box contains no direct semantics about the target and even results in ambiguity.

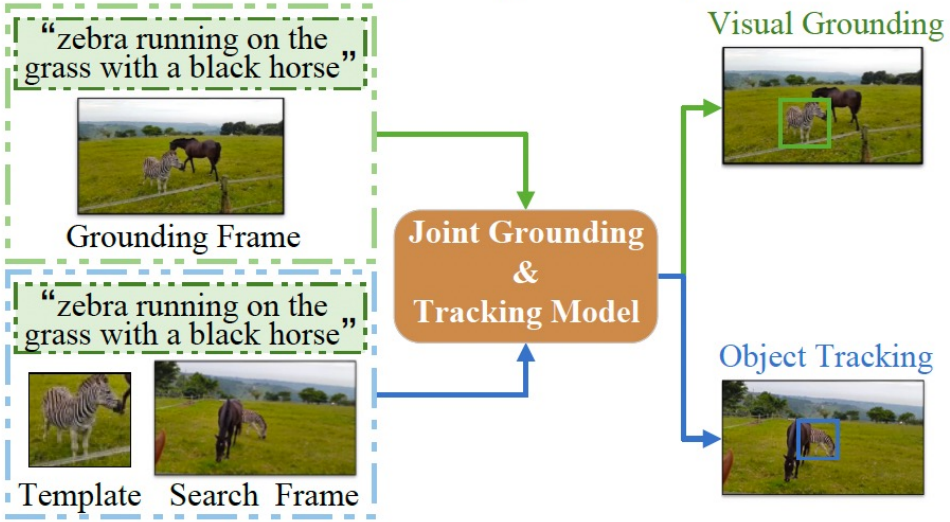
Wang X, Shu X, Zhang Z, et al. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. CVPR 2021 (pp. 13763-13773).

Background & Motivation

➢ Motivation:



(a) Separated visual grounding and tracking framework



(b) Joint visual grounding and tracking framework

Issues:

- Separated framework
- Not end-to-end training
- Ignore the natural language during tracking

Outline

1. Background & Motivation
2. Our proposed Methods
3. Experimental results
4. Visualization
5. Limitation and Conclusion

Our proposed Methods

➢ Framework:

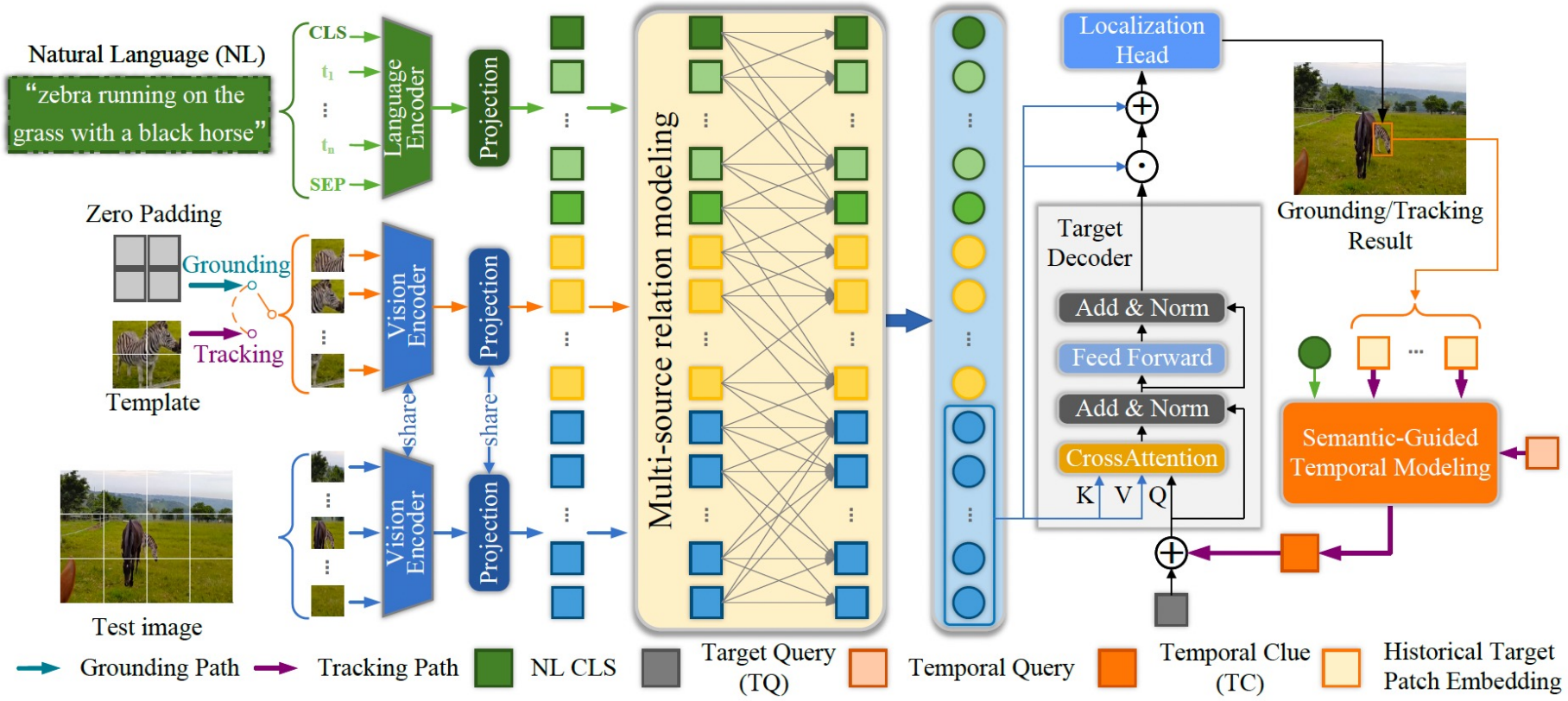


Figure 2. Overview of our joint visual grounding and tracking framework. Given a sequence and a natural language description, we first feed the description, the first frame (test image), and zero padding tokens into the model for visual grounding and accordingly obtain the template image. For each subsequent frame (test image), we feed it with the description and template image into the model for tracking. ⊙ and ⊕ denote the element-wise product and summation operations, respectively.

- **Integrates the processes of grounding and tracking by unifying the relation modeling of diverse references.**

Our proposed Methods

➢ Semantic-Guided Temporal Modeling:

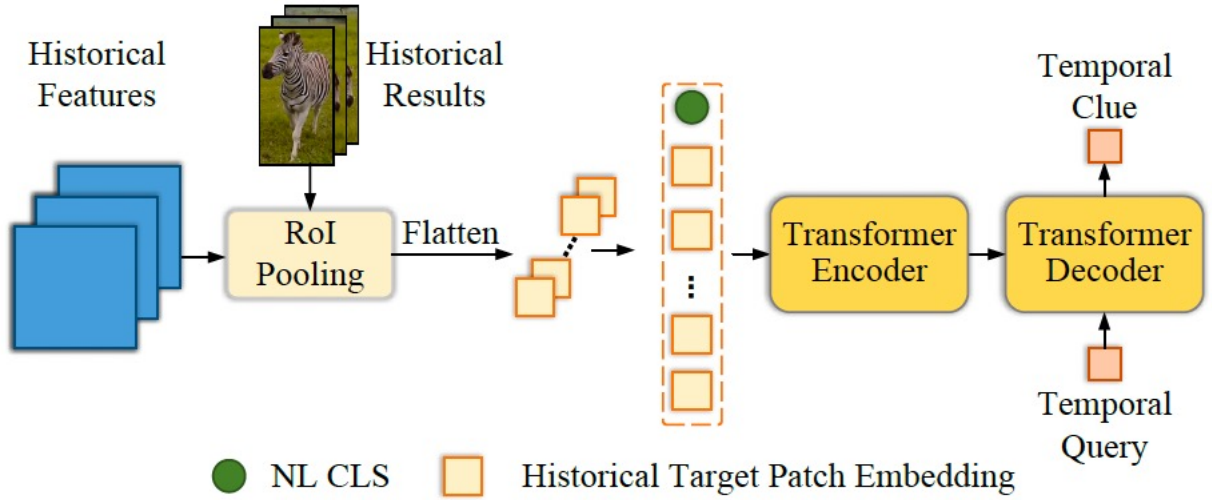


Figure 3. Architecture of the proposed semantic-guided temporal modeling module.

- Utilize the language context and historical target status
- Obtain the RoI feature from historical results

Outline

1. Background & Motivation
2. Our proposed Methods
3. Experimental results
4. Visualization
5. Limitation and Conclusion

Experimental results

➢ Performance on multi benchmarks:

Table 3. AUC and Precision (PRE) of different methods on the OTB99, LaSOT, and TNL2K datasets. The best and second-best results are marked in **bold** and underline. BB and NL denote the Bounding Box and Natural Language, respectively.

Algorithms	Initialize	OTB99		LaSOT		TNL2K	
		AUC	PRE	AUC	PRE	AUC	PRE
AutoMatch	BB	–		0.583	0.599	0.472	0.435
TrDiMP	BB	–		0.639	0.663	0.523	**0.528**
TransT	BB	–		0.649	0.690	0.507	0.517
STARK	BB	–		0.671	0.712	–	
KeepTrack	BB	–		0.671	0.702	–	
SwinTrack-B	BB	–		0.696	0.741	–	
OSTrack-384	BB	–		**0.711**	**0.776**	**0.559**	–
TNLS-II	NL	0.250	0.290	–		–	
RTTNLD	NL	0.540	**0.780**	0.280	0.280	–	
GTI	NL	0.581	0.732	0.478	0.476	–	
TNL2K-1	NL	0.190	0.240	0.510	0.490	0.110	0.060
CTRNLT	NL	0.530	0.720	0.520	0.510	0.140	0.090
Ours	NL	**0.592**	0.776	**0.569**	**0.593**	**0.546**	**0.550**
TNLS-III	NL+BB	0.550	0.720	–		–	
RTTNLD	NL+BB	0.610	0.790	0.350	0.350	0.250	0.270
TNL2K-2	NL+BB	0.680	0.880	0.510	0.550	0.420	0.420
SNLT	NL+BB	0.666	0.804	0.540	0.576	0.276	0.419
VLTTT	NL+BB	**0.764**	**0.931**	**0.673**	**0.721**	0.531	0.533
Ours	NL+BB	0.653	0.856	0.604	0.636	**0.569**	**0.581**

Experimental results

➢ Ablation Study:

Table 1. AUC and Precision (PRE) for four variants of our model on the LaSOT and TNL2K datasets.

Variants	LaSOT		TNL2K	
	AUC	PRE	AUC	PRE
SepRM	0.518	0.512	0.491	0.471
MSRM	0.536	0.550	0.511	0.500
MSRM-TDec	0.549	0.567	0.524	0.514
MSRM-TM	0.561	0.581	0.541	0.540
Our model	0.569	0.593	0.546	0.550

Table 2. Comparisons between separated and joint methods. We report FLOPs and inference time for grounding and tracking separately. Note that the time for all methods is tested on RTX 3090.

		Separated Model			Joint Model
		VLTVG+STARK	VLTVG+OSTrack	SepRM	Ours
FLOPs	Grounding	39.6G	39.6G	34.7G	34.9G
	Tracking	20.4G	48.3G	38.5G	42.0G
Time	Grounding	28.2ms	28.2ms	26.4ms	34.8ms
	Tracking	22.9ms	8.3ms	20.6ms	25.3ms
Params	Total	169.8M	214.7M	214.4M	153.0M
AUC	LaSOT	0.446	0.524	0.518	0.569
	TNL2K	0.373	0.399	0.491	0.546

➢ Performance on Visual Grounding:

Algorithms	NMTree [20]	LBYL-Net [14]	ReSC-Large [37]	TransVG [7]	VLTVG [36]	Ours
Accuracy	0.6178	0.6270	0.6312	0.6702	**0.7298**	<u>0.7007</u>

Outline

1. Background & Motivation
2. Our proposed Methods
3. Experimental results
4. Visualization
5. Limitation and Conclusion

➢ Visualization for revealing what temporal clues SGTM learn:

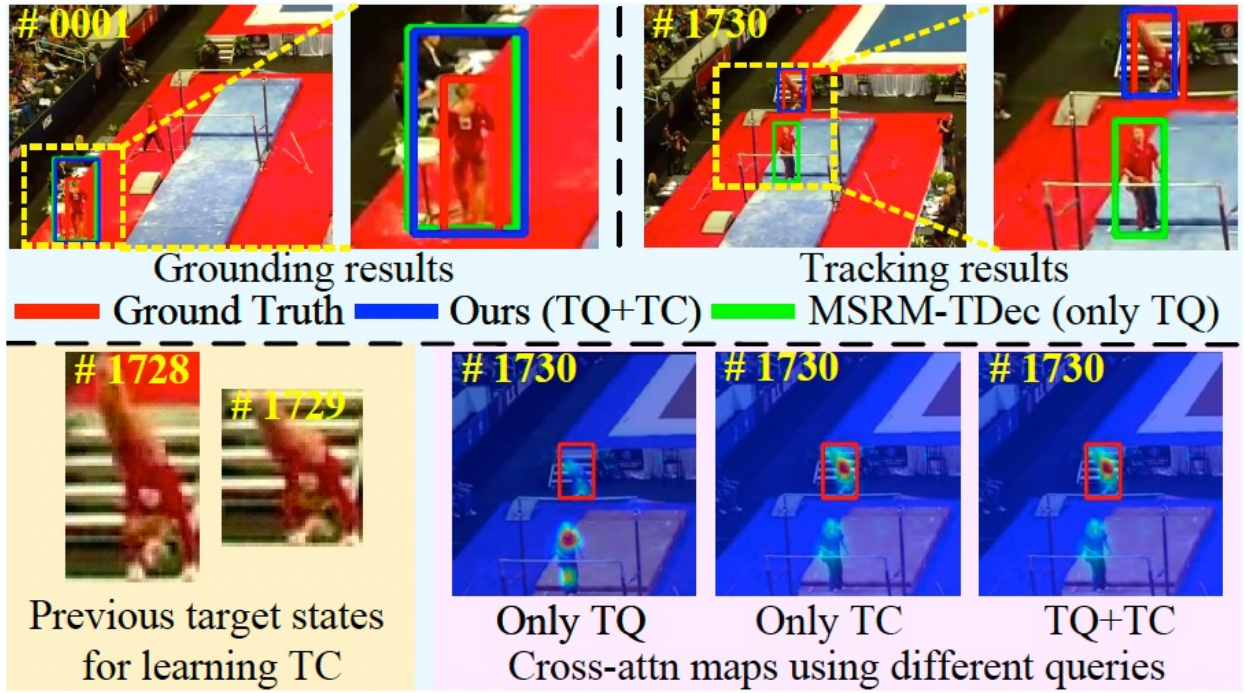


Figure 4. Visualization for revealing what temporal clue SGTM learns. "Only TQ" denotes only Target Query is used as the query in the target decoder. "Only TC" denotes only Temporal Clue is used as the query. "TQ+TC" denotes the summation of the Target Query and Temporal Clue is used as the query.

➢ Visualization for tracking results:

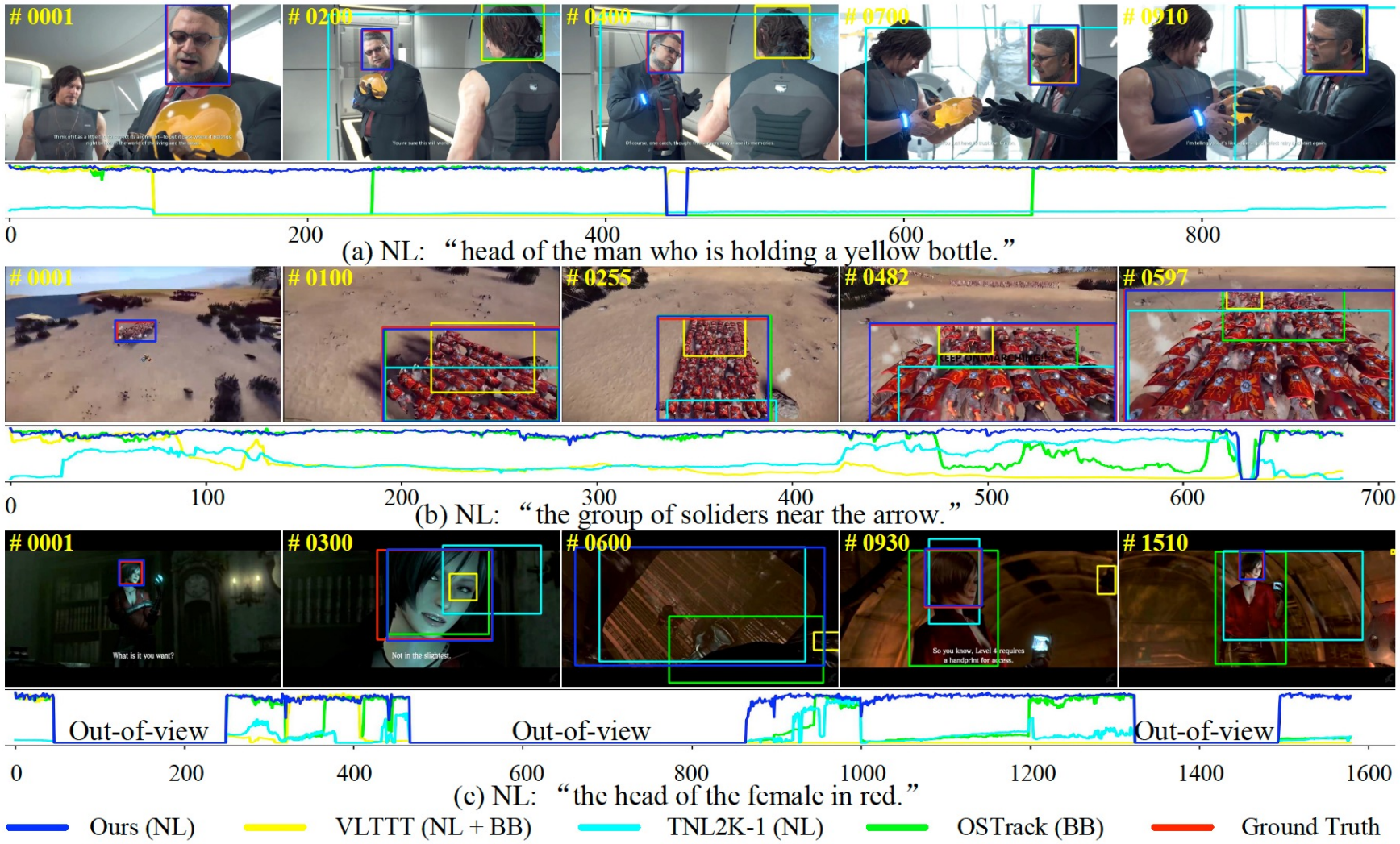
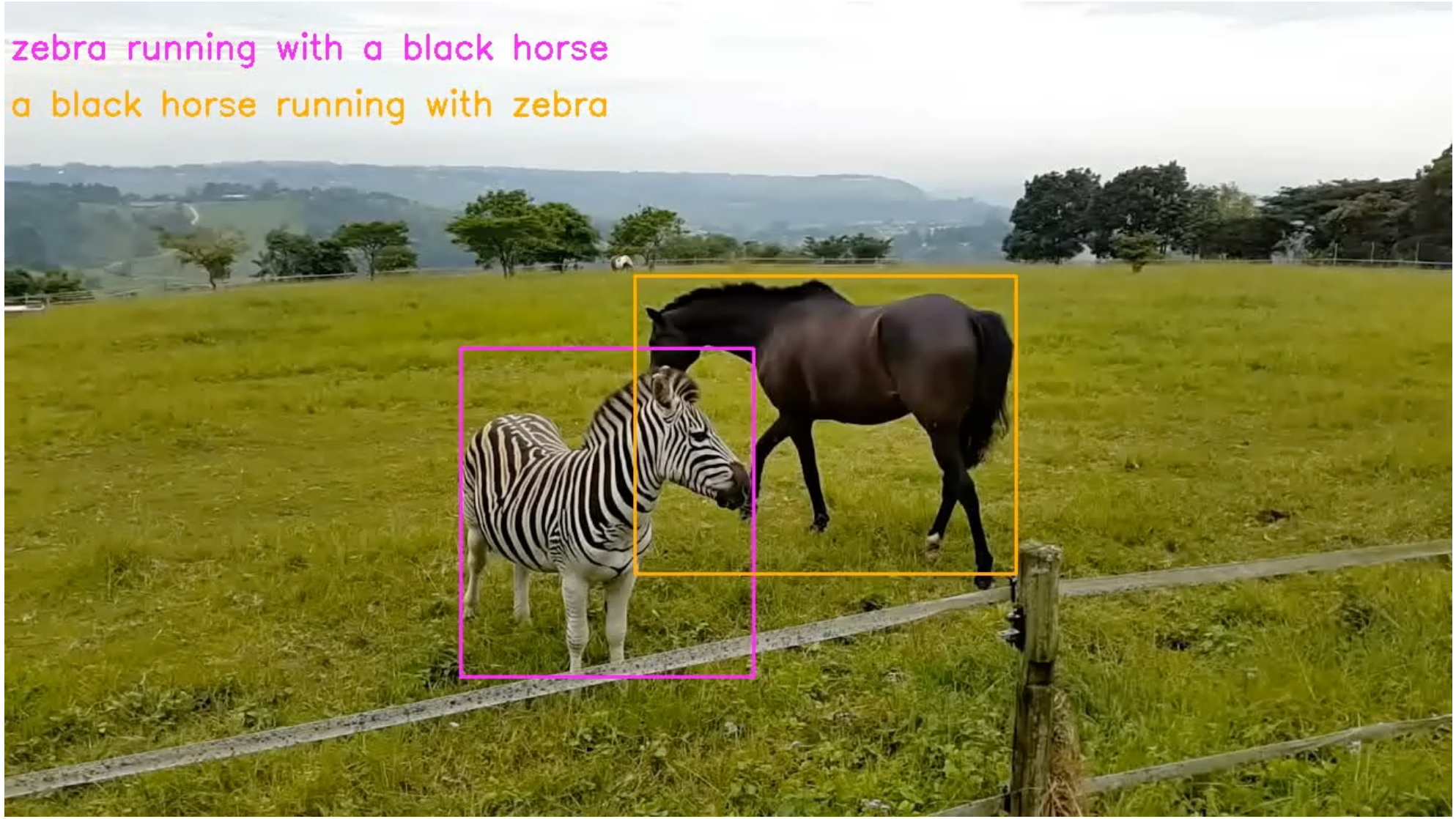


Figure 5. Qualitative comparison on three challenging sequences. From top to bottom, the main challenge factors are viewpoint change, appearance variation, and out-of-view, respectively. Our model is more robust than other trackers.

Visualization

➢ Visualization for a video:



Outline

1. Background & Motivation
2. Our proposed Methods
3. Experimental results
4. Visualization
5. Limitation and Conclusion

Limitation and Conclusion

➢ Limitation:

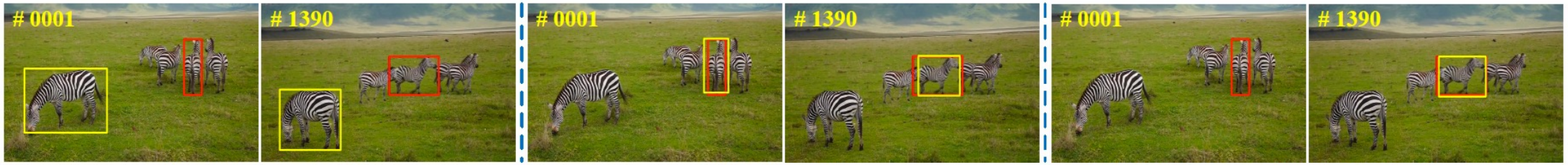


Figure 6. Analysis about the effect of ambiguous natural language (NL) on a zebra sequence. Given the original NL description (a) with ambiguity from LaSOT, our method localizes the wrong target at the first frame and consequently fails in the whole sequence. By contrast, given a clear NL description (b) or providing a bounding box (c) to eliminate ambiguity, our method can successfully locate the target.

- Sensitive to ambiguous natural language descriptions

Limitation and Conclusion

➢ Conclusion:

- We propose a joint visual grounding and tracking framework by unifying the relation modeling.

- We propose a semantics-guided temporal modeling module modeling the historical target states with global semantic information as guidance, which effectively improves tracking performance.

Thank for your listening