MambaVLT: Time-Evolving Multimodal State Space Model for Vision-Language Tracking

CVPR 2025 Highlight

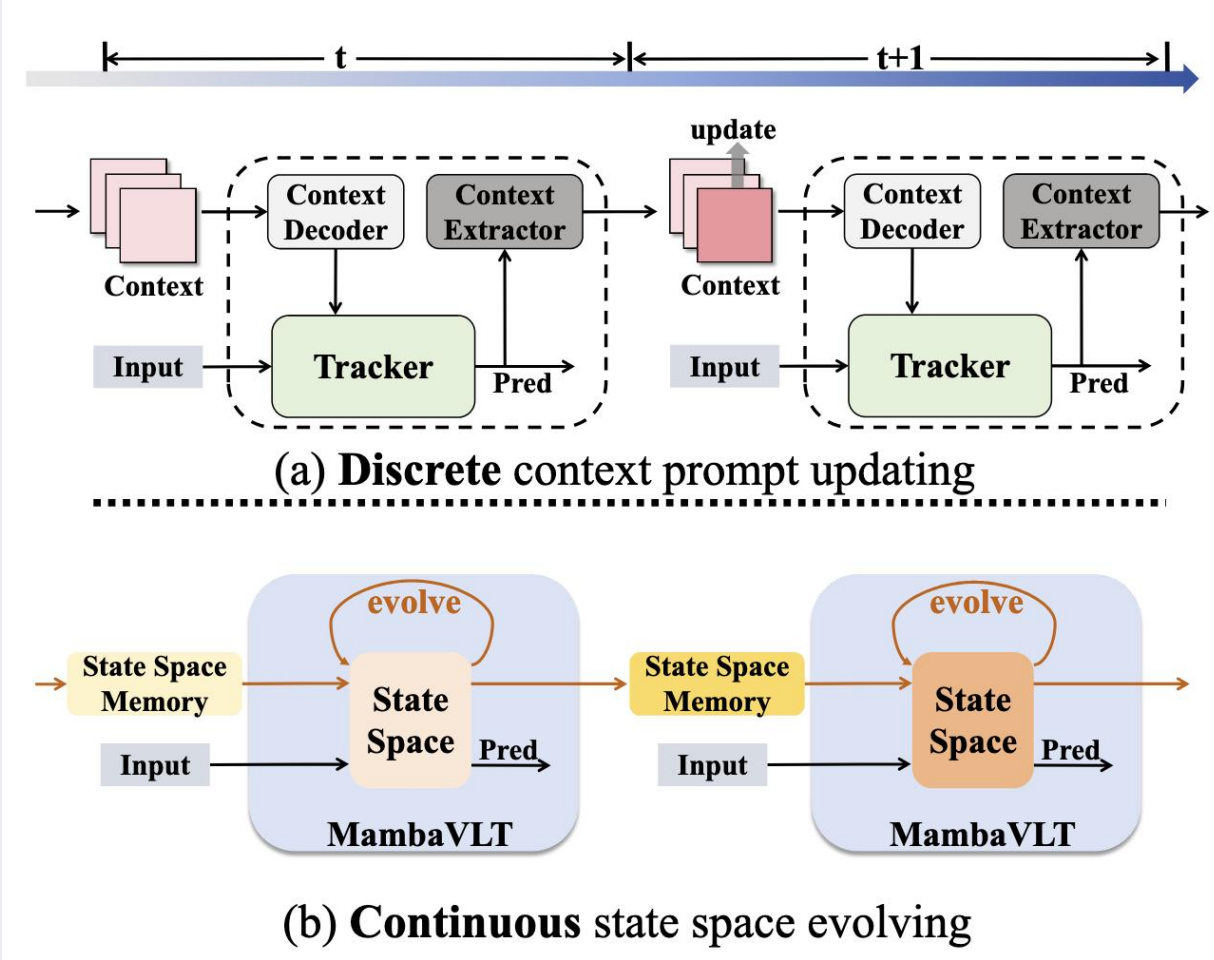
Xinqi Liu[1,†], Li Zhou[1,†], Zikun Zhou[2,*], Jianqiu Chen[1], and Zhenyu He[1,*]

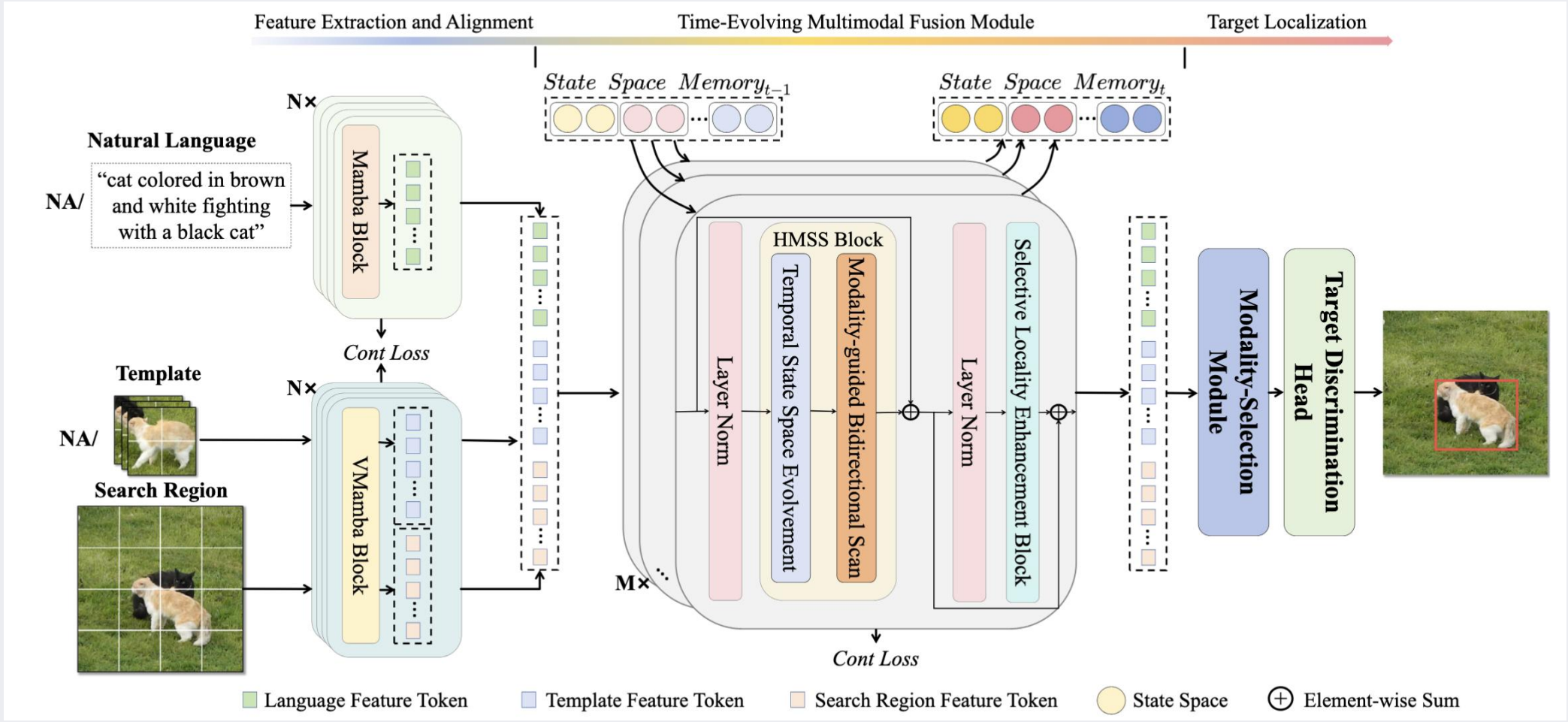
1 Harbin Institute of Technology, Shenzhen 2 Pengcheng Laboratory

†Equal Contribution *Corresponding Author

Motivation

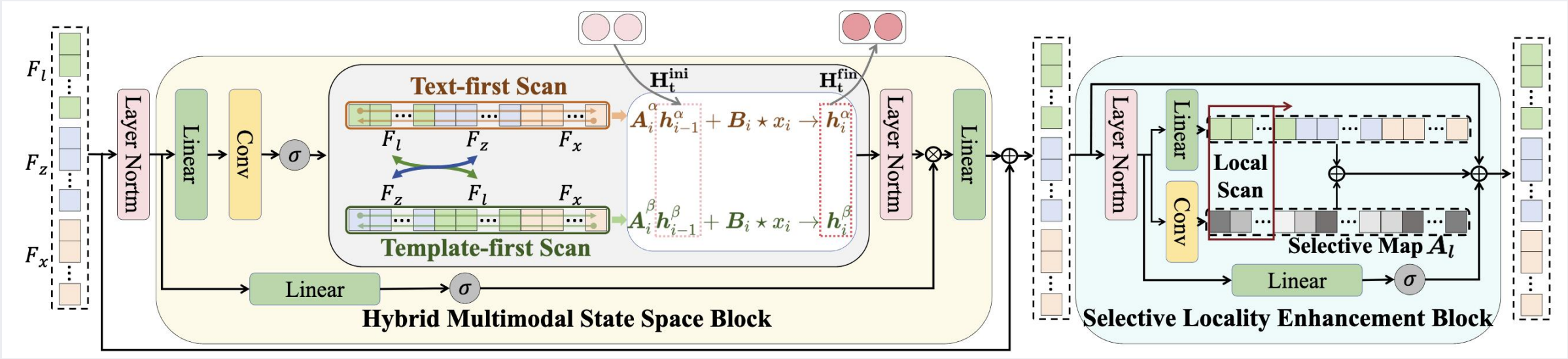
- Traditional vision-language trackers update references in **separate, discrete steps** which hinges on accurate predictions, causing error accumulation and an inability to fully exploit temporal cues. To overcome these limitations, we propose MambaVLT which adopts a **continuous, time-evolving state space mechanism** that can retain and update multimodal reference features in the video with linear complexity.





Method

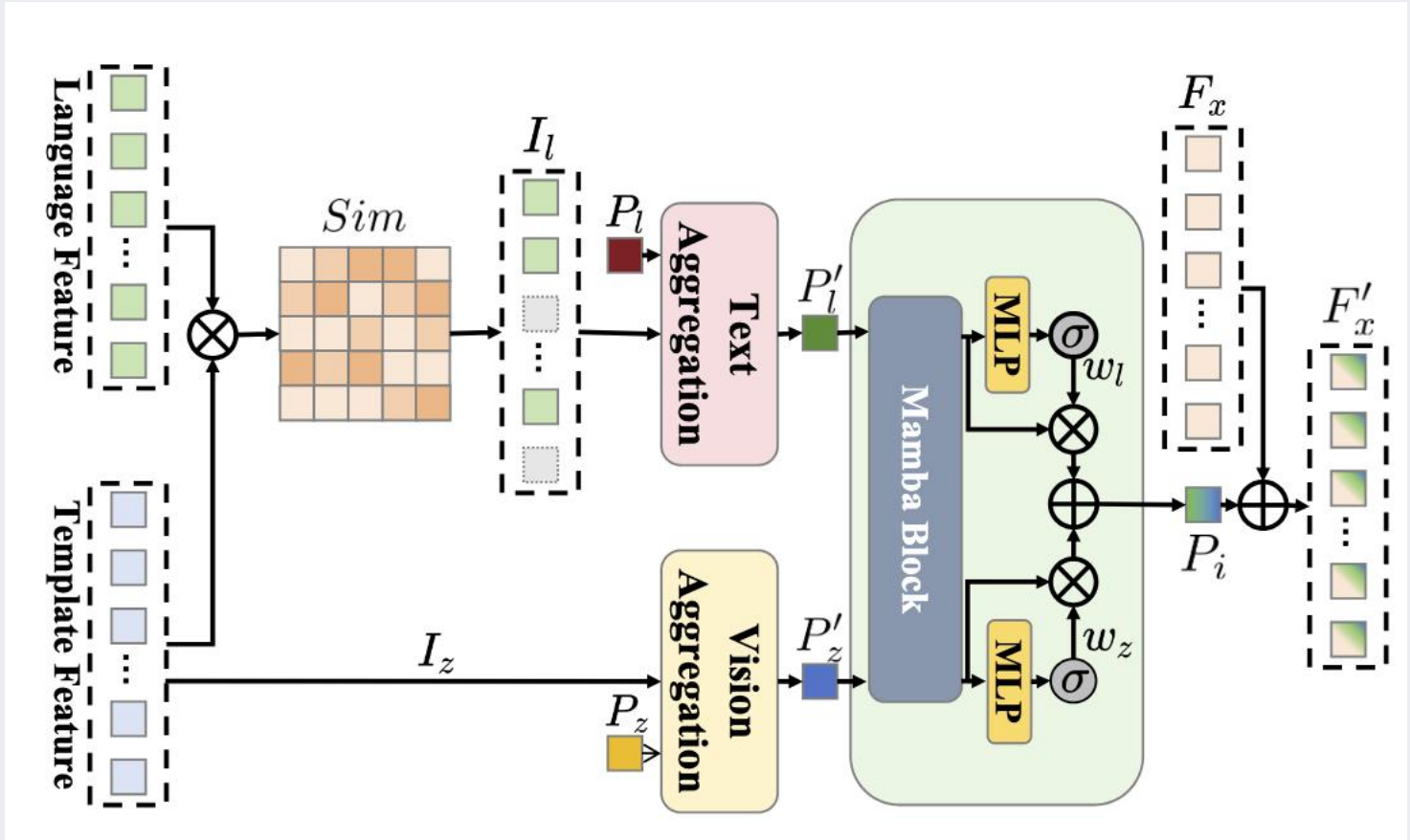
- Overview of the MambaVLT. Given various modality reference settings, features are initially extracted and aligned, then forwarded to the time-evolving multimodal fusion module. Subsequently, these features are input into the localization module to obtain precise localization information.



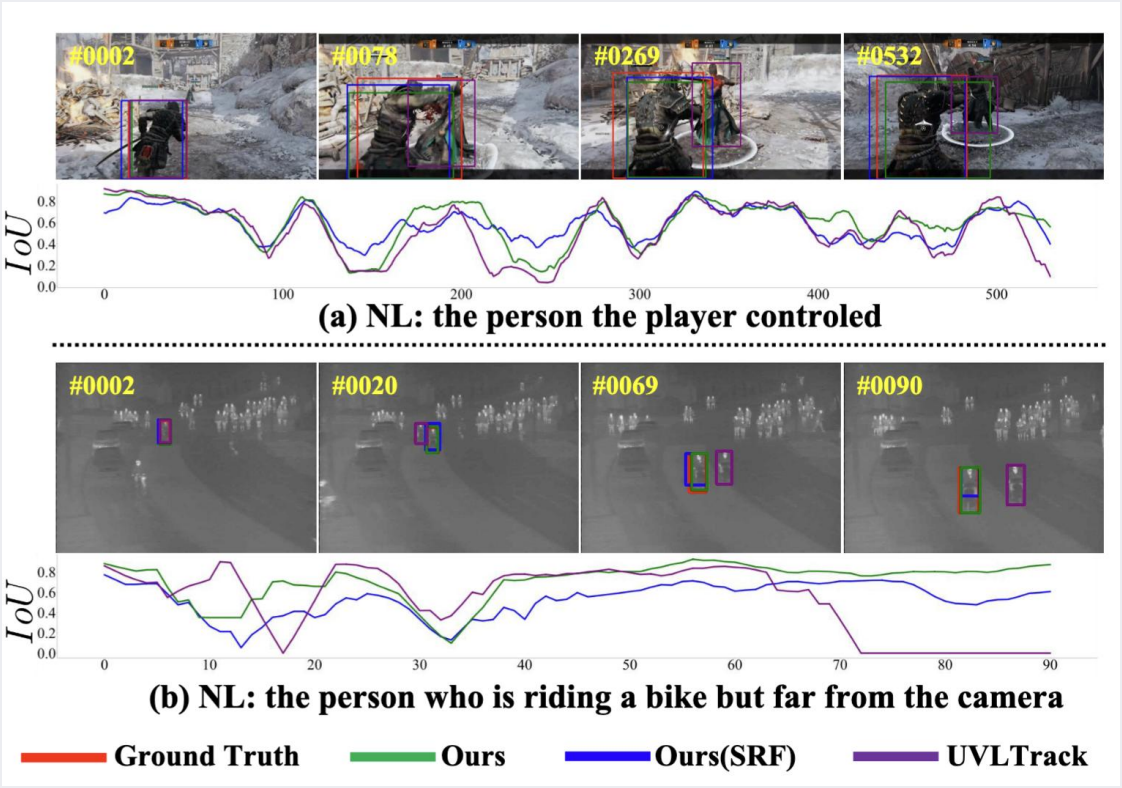
Method

- **Time-Evolving Multimodal Fusion**
  - **Hybrid State-Space Block** captures long-term temporal information by the time-evolving state space, based on which it models multimodal features and updates target reference information by a hybrid scan.
  - **Selective Locality Enhancement Block** performs a sliding window scan with a selective map to enhance multimodal features of the current time stamp through a global receptiveness.

Methods

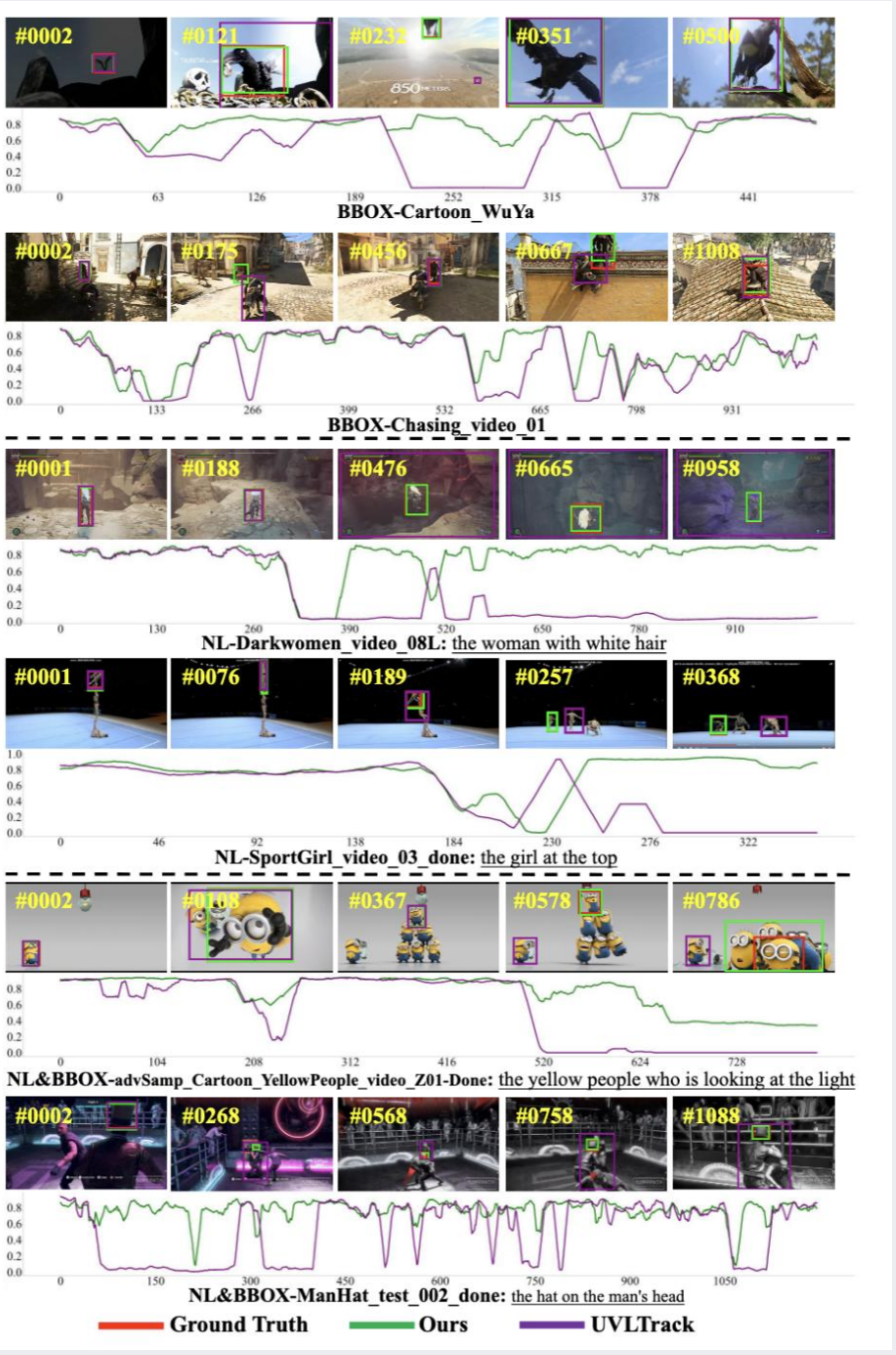


- In different tracking frames, the reliability of the language and template features may vary due to the target motion and appearance changes. Threfore, We further employ a **Modality-Selection Module** to selectively fuse multimodal reference features for search region feature refining.


#0002
#0078
#0269
#0532
IoU
(a) NL: the person the player controled
#0002
#0020
#0069
#0090
IoU
(b) NL: the person who is riding a bike but far from the camera
Ground Truth
Ours
Ours(SRF)
UVLTrack

Results


#0002
#0054
#0065
#0080 occlusion
#0002
occlusion
#0375
#0544
#0002
#0077
#0176 occlusion
#0355
Ground Truth
Ours
Ours (SRF)
UVLTrack


#0002
#0121
#0232
#0351
#0500
BBOX-Cartoon_WuYa
#0002
#0175
#0456
#0667
#1008
BBOX-Chasing_video_01
#0001
#0188
#0476
#0665
#0958
NL-Darkwomen_video_08L: the woman with white hair
#0001
#0076
#0189
#0257
#0368
NL-SportGirl_video_03_done: the girl at the top
#0002
#0367
#0578
#0786
NL&BBOX-advSamp_Cartoon_YellowPeople_video_Z01-Done: the yellow people who is looking at the light
#0002
#0268
#0568
#0758
#1088
NL&BBOX-ManHat_test_002_done: the hat on the man's head
Ground Truth
Ours
UVLTrack

哈爾濱工業大學(深圳)
HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

鹏城实验室
PENG CHENG LABORATORY

CVPR Nashville JUNE 11-15, 2025

Thanks for Your Watching!