UTPTrack: Towards Simple and Unified Token Pruning for Visual Tracking

Supplementary Material

To complement the manuscript, this appendix presents additional content and is organized as follows:

- Section A: Introduction of benchmarks.
- Section B: Practical efficiency.
- Section C: Overall performance (Sec. 4.2).
- Section D: Controlled-budget performance (Sec. 4.2).
- Section E: Additional ablation study and analysis.
- Section F: Visualization.

A. Introduction of Benchmarks

A.1. RGB-based Tracking Benchmarks

LaSOT. LaSOT [9] is a densely annotated, class-balanced large-scale long-term tracking benchmark. The benchmark consists of high-quality 1400 training set videos and 280 test set videos with an average video length of approximately 2500 frames, containing challenges of various complex scenarios. The evaluation metrics include Success (AUC) and Precision (P and P_{Norm}), among which AUC serves as the primary metric.

$LaSOT_{ext}$. $LaSOT_{ext}$ [10] extends the LaSOT benchmark to include 150 videos of 15 additional new object classes, which were carefully selected outside of ImageNet. The evaluation metrics remain consistent with LaSOT.

TrackingNet. TrackingNet [30] is a large-scale short-term tracking benchmark containing over 30K video sequences and over 14M bounding box annotations. The test set of TrackingNet contains 511 video sequences without publicly available ground truth. Researchers evaluate trackers by submitting their results to the official online evaluation server, which provides AUC, P and P_{Norm}.

GOT-10k. GOT-10K [16] is a challenging, high-diversity, large-scale benchmark with more than 10K video sequences, covering most of more than 560 categories of moving objects and more than 80 categories of motion patterns. The test set has 180 video sequences, including 84 object categories and 32 motion categories. It is worth noting that GOT-10k advocates a one-shot protocol, which means that the object classes between the train set and the test set are zero-overlapped. Similar to TrackingNet, the tracking results need to be submitted to the official evaluation server to obtain the tracker's performance. The evaluation metrics include Average Overlap (AO) and Success Rates at thresholds 0.5 and 0.75 ($SR_{0.5}$ and $SR_{0.75}$), among which AO serves as the primary metric.

A.2. Multimodal Tracking Benchmarks

VOT-RGBD22. VOT-RGBD22 [18] is a dataset designed for evaluating RGB-Depth (RGB-D) tracking algorithms, consisting of 127 video sequences. The evaluation follows the VOT protocol, including Accuracy, Robustness, and Expected Average Overlap (EAO), with EAO as the primary performance metric.

LasHeR. LasHeR [22] is a large-scale RGB-D dataset focused on long-term object tracking, containing 245 video sequences in its test set. The evaluation metrics include AUC and P, with the AUC as the primary metric.

RGBT234. RGBT234 [21] is an RGB-Thermal (RGB-T) tracking dataset, comprising 234 test video sequences that cover challenging scenarios such as illumination changes, occlusions, and low visibility. The evaluation metrics include Maximum Success Rate (MSR) and Maximum Precision Rate (MPR).

VisEvent. VisEvent [36] is a large-scale RGB-Event (RGB-E) object tracking dataset comprising 320 test video sequences. It emphasizes tracking performance under high-speed motion and low-light conditions. Researchers evaluate performance using the AUC and P metrics.

TNL2K. TNL2K [35] is a language-guided visual tracking dataset with 650 video sequences in the test set and covers a wide range of natural language descriptions. Researchers evaluate performance using AUC and P, with AUC as the primary metric.

OTB99. OTB99 [23] is a classic benchmark for short-term visual object tracking, with the test set consisting of 48 video sequences. Researchers assess tracking performance based on AUC and P scores.

B. Practical Efficiency Comparisons

As shown in Table 8, UTPTrack achieves higher FPS and lower per-layer backbone latency on both GPU and CPU across resolutions for OSTrack and SUTrack, while reducing training hours via lower backbone computation.

Table 8. Efficiency comparison. Lat.: per-layer backbone latency. GPU: NVIDIA 1080Ti; CPU: Intel Xeon Gold 6226R@2.90GHz.

Model	Train Hours	FPS (GPU)	FPS (CPU)	Lat. (GPU, ms)	Lat. (CPU, ms)
$OSTrack_{256}$	38.3	94.0	9.0	15.0	66.5
$UTPTrack\text{-}O_{256}$	32.6	95.0	16.7	12.7	49.5
$OSTrack_{384}$	101.4	39.8	3.2	36.1	340.8
$UTPTrack\text{-}O_{384}$	80.6	47.3	6.0	30.2	212.6
$SUTrack_{224}$	56.8	40.5	7.8	17.4	66.4
$UTPTrack\text{-}S_{224}$	53.3	42.7	9.3	17.8	56.4
$SUTrack_{384}$	209.2	27.4	3.4	63.3	340.2
$UTPTrack\text{-}S_{384}$	186.6	30.7	4.6	53.0	261.2

Table 9. SOTA comparison on RGB-based tracker across RGB-Based tracking.

Method	LaSOT			$\text{LaSOT}_{\text{ext}}$			TrackingNet			GOT-10k		
	AUC	P_{Norm}	P	AUC	P_{Norm}	P	AUC	P_{Norm}	P	AO	$\text{SR}_{0.5}$	$\text{SR}_{0.75}$
Trackers with Resolution 256												
OSTrack_{256} (Baseline)	67.9	76.9	73.3	47.6	57.5	53.6	84.5	89.0	83.0	74.9	84.2	73.0
OSTrack-CE_{256}	67.4	76.5	72.8	46.6	56.5	52.3	83.7	88.3	82.4	75.5	85.5	73.8
$\text{OSTrack-ToMe}_{256}$	67.6	76.6	72.6	46.5	56.0	52.5	**84.0**	88.6	82.6	73.4	83.4	72.4
$\text{OSTrack-EViT}_{256}$	**68.4**	**77.7**	**74.1**	46.1	55.7	51.1	**84.0**	88.7	**82.9**	**75.8**	**85.9**	**74.6**
UTPTrack-O_{256}(ours)	67.4	76.3	72.5	**47.3**	**57.1**	**53.3**	**84.0**	**88.8**	82.8	75.2	85.3	73.9
Trackers with Higher Resolution 384												
OSTrack_{384} (Baseline)	70.7	80.6	77.2	50.0	61.1	56.8	84.0	88.7	83.0	76.6	86.9	75.3
OSTrack-CE_{384}	70.6	80.1	**77.0**	50.6	61.2	57.0	84.1	88.8	83.0	76.3	86.5	74.9
$\text{OSTrack-ToMe}_{384}$	70.2	80.1	76.6	50.4	61.1	56.6	84.1	88.8	83.1	**77.1**	**87.2**	74.7
$\text{OSTrack-EViT}_{384}$	70.1	79.9	76.4	**52.1**	**62.9**	**59.2**	**84.3**	**89.0**	**83.3**	76.4	86.4	**75.3**
UTPTrack-O_{384}(ours)	**70.7**	**80.6**	**77.1**	49.4	59.8	55.8	84.0	88.7	82.7	76.5	86.9	74.5

Table 10. SOTA comparison on unified tracker across RGB-Based tracking.

Method	LaSOT			$\text{LaSOT}_{\text{ext}}$			TrackingNet			GOT-10k		
	AUC	P_{Norm}	P	AUC	P_{Norm}	P	AUC	P_{Norm}	P	AO	$\text{SR}_{0.5}$	$\text{SR}_{0.75}$
Trackers with Resolution 224												
SUTrack_{224} (Baseline)	73.7	83.8	80.7	53.2	64.5	61.6	85.9	90.4	85.7	77.5	86.7	78.1
SUTrack-CE_{224}	**72.8**	**82.9**	**80.2**	52.9	63.7	60.4	**85.3**	**90.0**	**84.8**	**77.6**	**87.6**	**78.0**
$\text{SUTrack-ToMe}_{224}$	71.4	81.2	78.3	52.0	62.9	59.3	85.2	**90.0**	84.6	77.5	87.4	77.5
$\text{SUTrack-EViT}_{224}$	71.3	81.9	78.3	51.3	62.7	58.5	84.5	89.5	83.5	76.4	86.5	75.7
$\text{SUTrack-DynamicViT}_{224}$	57.4	65.5	59.3	40.8	49.7	46.6	65.6	69.1	60.2	71.6	81.2	64.7
UTPTrack-S_{224}(ours)	72.6	82.8	79.8	**53.6**	**64.8**	**61.4**	85.2	**90.0**	84.7	77.3	87.1	77.7
Trackers with Higher Resolution 384												
SUTrack_{384} (Baseline)	73.9	83.2	81.2	52.8	63.3	60.3	86.8	90.9	87.3	79.1	87.6	80.2
SUTrack-CE_{384}	73.0	82.3	80.2	**54.1**	**64.7**	**61.9**	86.2	90.6	86.6	**79.7**	**89.0**	80.1
$\text{SUTrack-ToMe}_{384}$	74.2	83.8	**81.9**	53.0	63.3	60.5	86.3	90.7	86.8	79.1	88.1	79.1
$\text{SUTrack-EViT}_{384}$	73.7	83.4	81.6	53.0	63.7	61.0	86.1	90.6	86.4	78.4	87.9	78.1
$\text{SUTrack-DynamicViT}_{384}$	63.0	72.1	63.0	44.8	53.4	47.3	73.5	80.5	66.3	69.5	79.0	61.8
UTPTrack-S_{384}(ours)	**74.3**	**83.9**	81.8	53.6	64.3	61.5	**86.4**	**90.9**	**86.9**	79.3	88.3	**80.3**

C. Detailed Overall Performance Comparisons

In Sec. 4.2, we validate the effectiveness of UTPTrack under two tracking paradigms: RGB-based tracking and Unified tracking, which includes RGB-Depth (RGB-D), RGB-Thermal (RGB-T), RGB-Event (RGB-E), and RGB-Language (RGB-Lang) tasks. Using OSTrack and SUTrack as representative base models, we compare UTPTrack against state-of-the-art token compression methods, including CE, ToMe, EViT, and DynamicViT, across varying input resolutions. We report comprehensive results for Overall Performance, with detailed tabular data and analysis provided in this section.

RGB-based Tracking. We evaluate UTPTrack on four large-scale RGB-based tracking benchmarks, covering both long-term (LaSOT, $\text{LaSOT}_{\text{ext}}$) and short-term (TrackingNet, GOT-10k) scenarios. The results are summarized in Tab. 9 and Tab. 10. For the RGB-based variant, UTPTrack-O maintains the accuracy of base model across all four benchmarks and both resolutions. At 256 resolution, UTPTrack-O_{256} closely matches base model on LaSOT, $\text{LaSOT}_{\text{ext}}$, and TrackingNet while slightly improving AO on GOT-10k, and remains competitive with other compression methods. At 384 resolution, UTPTrack-O_{384} similarly matches the performance of base model across all benchmarks. In con-

Table 11. SOTA comparison on unified tracker across RGB-Depth, RGB-Thermal, RGB-Event, and RGB-Language tracking.

Method	VOT-RGBD22			LasHeR		RGBT234		VisEvent		TNL2K		OTB99	
	EAO	Acc	Rob	AUC	P	MSR	MPR	AUC	P	AUC	P	AUC	P
Trackers with Resolution 224													
SUTrack_{224} (Baseline)	75.5	82.5	91.3	59.9	74.8	70.0	92.1	63.3	80.7	67.8	73.8	68.4	91.1
SUTrack-CE_{224}	**76.7**	**82.7**	**92.4**	**59.6**	**74.0**	70.1	92.4	**62.8**	79.9	65.7	71.1	70.8	92.7
$\text{SUTrack-ToMe}_{224}$	76.1	**82.7**	91.8	59.2	73.7	69.4	91.8	61.7	78.9	**65.9**	**71.4**	70.5	91.6
$\text{SUTrack-EViT}_{224}$	75.0	82.1	90.9	58.6	73.0	69.2	91.6	62.0	79.4	65.0	70.0	70.8	91.9
$\text{SUTrack-DynamicViT}_{224}$	62.8	74.2	84.4	46.1	57.5	63.2	85.1	50.9	71.8	55.6	55.0	65.8	86.8
UTPTrack-S_{224}(ours)	76.4	**82.7**	92.1	59.2	73.9	**70.4**	**92.9**	62.6	**80.0**	65.6	71.0	**71.9**	**93.7**
Trackers with Higher Resolution 384													
SUTrack_{384} (Baseline)	76.0	83.2	91.5	60.6	75.3	69.2	92.4	63.0	79.3	68.2	74.7	69.2	91.0
SUTrack-CE_{384}	**77.9**	83.6	**92.8**	**60.4**	**75.0**	**69.7**	92.0	**62.9**	79.5	66.2	72.6	70.0	90.6
$\text{SUTrack-ToMe}_{384}$	76.8	**83.8**	91.5	59.7	74.1	69.4	92.0	62.6	79.3	66.4	72.8	70.8	91.8
$\text{SUTrack-EViT}_{384}$	77.1	83.2	92.5	58.8	73.2	68.2	90.8	62.1	79.2	66.2	72.7	**72.5**	**94.6**
$\text{SUTrack-DynamicViT}_{384}$	71.2	81.0	87.8	51.6	63.5	56.7	82.2	56.4	73.4	58.9	58.0	68.0	86.5
UTPTrack-S_{384}(ours)	76.8	**83.8**	91.6	**60.4**	74.8	**69.7**	**92.2**	62.8	**79.7**	**66.6**	**72.9**	72.2	93.7

trast, the unified variant, UTPTrack-S, consistently delivers strong results across all benchmarks. UTPTrack-S_{224} improves the AUC on $\text{LaSOT}_{\text{ext}}$ and TrackingNet over base model, while UTPTrack-S_{384} achieves the best or second-best scores on all four datasets. These results demonstrate that our pruning strategy effectively preserves accuracy across different resolutions and tracking paradigms.

RGB-D/T/E Tracking. As shown in Tab.11, UTPTrack-S_{224} and UTPTrack-S_{384} consistently rank among the top two across all benchmarks, including VOT-RGBD22 (RGB-D), LasHeR and RGBT234 (RGB-T), and VisEvent (RGB-E). This demonstrates the robustness and generalizability of our pruning strategy across diverse modalities and dataset scales, while maintaining strong performance and high efficiency.

RGB-Lang Tracking. As shown in Tab. 11, UTPTrack-S significantly outperforms the baseline model on the short-term tracking dataset OTB99 across different resolutions. It achieves gains of 3.5% at 224 resolution and 2.0% at 384 resolution. On TNL2K, however, the performance of UTPTrack-S decreases relative to the baseline model, likely due to compression-induced performance loss. Even so, UTPTrack-S still delivers competitive performance on TNL2K, and at 384 resolution it achieves the best results among all compared methods. Such degradation across all compressed models may be attributed to the limited precision of the initial language descriptions in later stages of long videos, especially in long-term sequences, which reduces the reliability of language-guided cues.

D. Detaild Controlled-Budget Performance Comparisons

In Sec. 4.2, we conduct controlled-budget experiments under fixed compression ratios to ensure fair comparisons for RGB-based and unified tracking, with comprehensive results presented in this section.

RGB-based Tracking. For RGB-based tracking, we compare UTPTrack with other methods on four widely used RGB-based benchmarks. As shown in Tab. 12, UTPTrack-O consistently outperforms all other methods across all pruning rates. Notably, with 18.8% of the visual tokens pruned, UTPTrack-O_{256} outperforms the baseline model by 0.2%, suggesting that removing redundant tokens can even lead to performance gains. Similarly, in Tab. 16, UTPTrack-S surpasses all other compression methods across all pruning rates, and at a retained ratio of 35.4% visual tokens, it achieves a 1.0% improvement over the second-best method.

RGB-D/T/E Tracking. As shown in Tab. 17, UTPTrack outperforms the baseline on both the RGB-D and RGB-T benchmarks across different pruning ratios. On the VOT-RGBD22 dataset, it surpasses the baseline by 1.3% when pruning 25.6% visual tokens. On the RGB-E benchmark, UTPTrack consistently ranks among the top two methods. These results demonstrate that UTPTrack generalizes effectively across diverse multimodal tracking scenarios.

RGB-Lang Tracking. As shown in Tab. 17, UTPTrack maintains competitive performance across different compression rates on the RGB-Language benchmark. At all compression rates, UTPTrack outperforms the baseline on

Table 12. Performance comparisons under different vision token compression settings for **RGB-based tracking**. The baseline uses 384 tokens. The first row shows raw scores, and the second row reports percentages relative to the upper bound.

Method	LaSOT			$\text{LaSOT}_{\text{ext}}$			TrackingNet			GOT-10k			Avg Perf.(%)
	AUC	P_{Norm}	P	AUC	P_{Norm}	P	AUC	P_{Norm}	P	AO	$\text{SR}_{0.5}$	$\text{SR}_{0.75}$	
Upper Bound, 384 Tokens (100%)													
OSTrack_{256} (Baseline)	67.9 100%	76.9 100%	73.3 100%	47.6 100%	57.5 100%	53.6 100%	84.5 100%	89.0 100%	83.0 100%	74.9 100%	84.2 100%	73.0 100%	100%
Retain 87.2% Tokens (≈ 335 tokens, ↓ 18.8%)													
OSTrack-CE_{256}	67.1 98.9%	75.8 98.6%	72.0 98.3%	46.9 98.6%	56.7 98.6%	52.4 97.7%	84.1 99.6%	88.8 99.8%	83.0 99.9%	74.9 100.0%	85.0 101.0%	73.5 100.7%	<u>99.3%</u>(↓ 0.7%)
$\text{OSTrack-ToMe}_{256}$	68.0 100.2%	76.9 99.9%	72.9 99.5%	46.4 97.5%	56.3 98.0%	51.5 96.1%	83.8 99.3%	88.7 99.6%	82.7 99.6%	74.3 99.2%	84.2 100.0%	73.0 100.0%	99.0%(↓ 1.0%)
$\text{OSTrack-EViT}_{256}$	66.8 98.4%	75.7 98.4%	71.7 97.9%	46.7 98.1%	56.3 98.1%	52.3 97.5%	84.1 99.5%	88.9 99.8%	82.8 99.7%	75.0 100.1%	85.2 101.2%	73.9 101.2%	99.0%(↓ 1.0%)
UTPTrack-O_{256}	67.1 98.9%	75.7 98.5%	72.2 98.5%	48.3 101.5%	58.3 101.5%	54.4 101.4%	84.0 99.5%	88.7 99.6%	83.0 99.9%	75.6 100.9%	85.8 101.9%	74.7 102.3%	**100.2%**(↑ 0.2%)
Retain 75.5% Tokens (≈ 290 tokens, ↓ 24.5%)													
OSTrack-CE_{256}	67.3 99.1%	76.2 99.1%	72.6 99.1%	47.3 99.4%	57.2 99.5%	53.1 99.0%	84.2 99.7%	89.0 100.0%	83.2 100.2%	74.2 99.1%	84.1 99.9%	72.8 99.7%	99.3%(↓ 0.7%)
$\text{OSTrack-ToMe}_{256}$	68.2 100.4%	77.1 100.2%	73.2 100.0%	46.5 97.6%	56.0 97.5%	52.2 97.4%	84.2 99.7%	89.0 100.0%	83.2 100.1%	74.9 100.0%	85.1 101.1%	73.6 100.8%	<u>99.4%</u>(↓ 0.6%)
$\text{OSTrack-EViT}_{256}$	67.8 99.8%	76.8 99.9%	73.4 100.2%	47.1 99.0%	56.6 98.6%	52.8 98.4%	83.8 99.2%	88.5 99.4%	82.5 99.4%	72.9 97.3%	82.8 98.3%	70.4 96.4%	98.8%(↓ 1.2%)
UTPTrack-O_{256}	67.6 99.5%	76.3 99.3%	72.9 99.5%	47.3 99.5%	57.2 99.6%	53.0 98.9%	83.8 99.2%	88.4 99.3%	82.7 99.6%	74.8 99.9%	85.1 101.1%	73.5 100.7%	**99.5%**(↓ 0.5%)
Retain 65.6% Tokens (≈ 252 tokens, ↓ 34.4%)													
OSTrack-CE_{256}	67.2 99.0%	76.1 99.0%	72.5 98.9%	46.5 97.8%	56.3 97.9%	51.8 96.6%	83.3 98.6%	88.2 99.0%	82.0 98.7%	72.8 97.2%	82.9 98.5%	70.9 97.1%	<u>98.1%</u>(↓ 1.9%)
$\text{OSTrack-ToMe}_{256}$	66.3 97.7%	75.7 98.4%	71.1 97.0%	46.0 96.7%	55.9 97.2%	50.9 94.9%	82.5 97.7%	87.5 98.2%	80.2 96.6%	71.0 94.8%	81.3 96.6%	67.9 93.0%	96.7%(↓ 3.3%)
$\text{OSTrack-EViT}_{256}$	68.1 100.3%	77.2 100.4%	73.2 99.9%	46.1 96.8%	55.8 97.1%	51.2 95.5%	83.0 98.2%	87.9 98.7%	81.7 98.4%	72.8 97.2%	83.3 98.9%	69.7 95.5%	<u>98.1%</u>(↓ 1.9%)
UTPTrack-O_{256}	68.2 100.4%	77.3 100.5%	74.0 101.1%	46.2 97.0%	56.1 97.7%	51.8 96.6%	84.0 99.5%	88.9 99.8%	83.1 100.1%	74.9 100.0%	85.0 101.0%	73.3 100.4%	**99.2%**(↓ 0.8%)

the short-term tracking dataset OTB99. With 25.6% visual tokens pruned, UTPTrack surpasses the baseline by 3.3%, and even at a pruning rate of 64.6%, it still leads the baseline by 2.7%. On TNL2K, UTPTrack consistently ranks in the top two compared to other compression methods.

E. Additional Ablation Study

E.1. The Effect of Pruning Location for RGB-based Tracking

To analyze the impact of pruning locations, we systematically vary the pruning positions of both the CE and DTE modules across different transformer layers, with detailed results summarized in Tab. 13. Based on empirical results and a meticulous consideration of the performance-efficiency trade-off, we select the pruning configuration (#3) that prunes CE at layers [3, 6, 9], and DTE at layers [4, 7, 10]. This arrangement achieves an optimal balance, maintaining competitive tracking performance while substantially reducing the computational overhead.

Table 13. Ablation Study on CTEM Location for **RGB-based Trackers** with a 12-layer ViT backbone.

#	CE Location	DTE Location	LaSOT	$\text{LaSOT}_{\text{ext}}$	TrackingNet	GOT-10k	Avg Vis Tok	Cm Vis Tok	MACs (G)	Avg Perf. (%)
1	–	–	67.9	47.6	84.5	74.9	384.0	384	34.5	100.0
2	[3,6,9]	[3,6,9]	67.6	47.0	83.8	74.5	267.8	176	25.1	99.3
3	[3,6,9]	[4,7,10]	68.4	46.6	83.8	75.3	271.2	176	25.4	**99.6**
4	[3,6,9]	[2,5,8]	67.4	45.7	83.8	75.4	264.3	176	24.8	98.8
5	[2,5,8]	[3,6,9]	68.0	46.4	83.9	75.7	253.8	176	23.9	<u>99.5</u>
6	[4,7,10]	[3,6,9]	67.5	46.5	83.9	73.4	281.7	176	26.3	98.6

E.2. The Effect of Spatial Priors in Token Type-Aware Strategy

TTA does not replace attention-based importance estimation, but stabilizes static template pruning, which is sensitive due to limited tokens and foreground–background imbalance. The bounding-box prior uses only initialization information. As shown in Table 14, attention-based pruning without priors consistently outperforms random pruning, while adding spatial priors yields modest but consistent gains. These results indicate that TTA complements attention-based pruning by improving stability under high compression ratios.

Table 14. Effect of spatial priors in attention-guided pruning.

Method	LaSOT	$\text{LaSOT}_{\text{ext}}$	TrackingNet	GOT-10k
Attn-based Pruning w/o Priors	67.2	46.5	**84.3**	74.4
Attn-based Pruning w/ Priors	**67.4**	**47.3**	84.0	**75.2**
Random Pruning w/ Priors	67.0	46.8	83.9	74.2

E.3. The Effect of Token Type-Aware Strategy for Unified Tracking

Table 15 compares the effects of three foreground bonus strategies on ST pruning for unified trackers, corresponding to the Full, Soft and All variants introduced in Sec. 3.2.1. Among them, the Soft bonus delivers the best overall performance across the five benchmarks, achieving an average accuracy of 99.8%, outperforming the Full and All strategies by approximately 0.5% and 0.7%, respectively. Under a strict pruning ratio, the Full bonus rewards only patches that lie entirely within the target box, making it overly conservative and prone to underestimating foreground tokens near object boundaries. Conversely, the All bonus assigns the maximum reward as long as a patch merely intersects the box, which is excessively permissive and easily introduces background noise. The Soft bonus, computed as the mean mask value within each patch, provides a more fine-grained estimate of foreground coverage and yields smoother boundary transitions, making it a more reliable continuous signal for guiding pruning.

Table 15. Ablation Study on bonus for Unified Trackers.

#	Bonus	LaSOT	VOT-RGBD22	LasHeR	VisEvent	OTB99	Avg Perf. (%)
1	Full	71.7	75.7	59.5	61.7	69.7	99.3
2	Soft	72.4	75.9	59.4	61.9	70.7	**99.8**
3	All	72.2	75.6	59.5	61.2	69.4	99.1

F. Visualization

We present visualizations of the UTPTrack pruning process for RGB-based, RGB-Lang, RGB-D, RGB-T, and RGB-E tracking, as shown in Figs. 5 to 9. The top-left shows the search region image, with the static template on the bottom-left and the dynamic template on the bottom-right. Masked areas indicate pruned tokens. The six stages illustrate a progressive schedule: stages 1, 3, and 5 progressively prune tokens from the search region, while stages 2, 4, and 6 further prune tokens from both the static and dynamic templates.

Table 16. Performance comparisons under different vision token compression configurations across **unified tracking** on RGB-based tracking benchmark. The vanilla number of visual tokens is 294. The first line of each method is the raw performance of benchmarks, and the second line is the proportion relative to the upper limit. The average performance listed is calculated across all 10 benchmarks.

Method	LaSOT			$\text{LaSOT}_{\text{ext}}$			TrackingNet			GOT-10k			Avg Perf.(%)
	AUC	P_{Norm}	P	AUC	P_{Norm}	P	AUC	P_{Norm}	P	AO	$\text{SR}_{0.5}$	$\text{SR}_{0.75}$	
Upper Bound, 294 Tokens (100%)													
SUTrack_{224} (Baseline)	73.7 100%	83.8 100%	80.7 100%	53.2 100%	64.5 100%	61.6 100%	85.9 100%	90.4 100%	85.7 100%	77.5 100%	86.7 100%	78.1 100%	100%
Retain 71.4% Tokens (≈ 218 tokens, ↓ 25.6%)													
SUTrack-CE_{224}	72.4 98.3%	82.4 98.4%	79.2 98.1%	52.8 99.1%	63.6 98.7%	60.2 97.7%	85.1 99.0%	89.7 99.2%	84.4 98.5%	77.4 99.9%	87.0 100.3%	78.0 99.9%	<u>99.5</u>%(↓ 0.5%)
$\text{SUTrack-ToMe}_{224}$	72.1 97.8%	82.0 97.8%	78.9 97.7%	52.6 98.8%	63.7 98.8%	59.7 96.9%	85.4 99.4%	90.0 99.6%	84.7 98.9%	76.9 99.2%	86.5 99.8%	77.2 98.8%	99.4%(↓ 0.6%)
$\text{SUTrack-EViT}_{224}$	71.6 97.1%	81.9 97.7%	79.0 97.8%	52.3 98.3%	63.2 98.1%	59.5 96.6%	84.9 98.8%	89.7 99.2%	84.1 98.2%	76.9 99.2%	87.2 100.6%	76.8 98.3%	99.0%(↓ 1.0%)
$\text{SUTrack-DyViT}_{224}$	68.8 93.4%	81.6 97.4%	71.1 88.0%	49.7 93.4%	63.0 97.8%	55.3 89.8%	80.2 93.4%	86.4 95.6%	74.9 87.4%	74.6 96.3%	87.1 100.5%	72.1 92.3%	96.5%(↓ 3.5%)
UTPTrack-S_{224}	72.1 97.9%	82.2 98.1%	78.9 97.7%	52.9 99.4%	63.8 98.9%	60.1 97.5%	85.1 99.0%	89.7 99.3%	84.5 98.6%	77.7 100.3%	87.4 100.8%	78.2 100.1%	**99.8%**(↓ 0.2%)
Retain 52.0% Tokens (≈ 153 tokens, ↓ 48.0%)													
SUTrack-CE_{224}	72.0 97.7%	81.8 97.6%	78.9 97.7%	52.8 99.2%	63.9 99.1%	60.4 97.9%	85.3 99.3%	90.0 99.6%	84.7 98.8%	77.1 99.5%	86.8 100.1%	77.7 99.5%	<u>99.2</u>%(↓ 0.8%)
$\text{SUTrack-ToMe}_{224}$	71.4 96.9%	81.2 96.9%	78.3 96.9%	52.0 97.8%	62.9 97.6%	59.3 96.2%	85.2 99.1%	90.0 99.6%	84.6 98.7%	77.5 100.0%	87.4 100.8%	77.5 99.2%	99.0%(↓ 1.0%)
$\text{SUTrack-EViT}_{224}$	71.2 96.6%	81.6 97.4%	78.4 97.1%	51.9 97.5%	61.1 94.8%	58.9 95.6%	84.7 98.6%	89.7 99.3%	84.0 98.0%	76.9 99.2%	87.2 100.6%	76.9 98.5%	98.5%(↓ 1.5%)
$\text{SUTrack-DyViT}_{224}$	53.2 72.1%	60.6 72.3%	55.3 68.5%	36.5 68.5%	44.1 68.4%	41.7 67.6%	58.2 67.7%	62.4 69.0%	52.2 61.0%	52.2 67.4%	60.0 69.2%	34.8 44.6%	78.8%(↓ 21.2%)
UTPTrack-S_{224}	72.5 98.4%	82.5 98.5%	79.6 98.6%	53.1 99.7%	64.0 99.3%	60.5 98.2%	85.5 99.5%	90.1 99.7%	84.9 99.1%	77.1 99.5%	86.8 100.1%	77.5 99.2%	**99.5**%(↓ 0.5%)
Retain 35.4% Tokens (≈ 104 tokens, ↓ 64.6%)													
SUTrack-CE_{224}	70.5 95.7%	80.5 96.1%	77.2 95.6%	51.8 97.3%	63.0 97.7%	59.3 96.2%	84.8 98.7%	89.5 99.1%	83.9 97.9%	75.9 97.9%	85.9 99.1%	75.3 96.4%	<u>98.3</u>%(↓ 1.7%)
$\text{SUTrack-ToMe}_{224}$	67.9 92.1%	73.9 88.1%	70.7 87.6%	48.6 91.3%	55.9 86.6%	53.8 87.2%	80.0 93.1%	83.1 92.0%	75.5 88.1%	71.6 92.4%	81.2 93.7%	64.7 82.8%	92.5%(↓ 7.5%)
$\text{SUTrack-EViT}_{224}$	69.6 94.5%	80.0 95.4%	76.5 94.7%	50.6 95.1%	61.8 95.9%	57.6 93.4%	84.3 98.1%	89.2 98.8%	83.2 97.0%	74.0 95.5%	84.3 97.2%	72.4 92.7%	96.7%(↓ 3.3%)
$\text{SUTrack-DyViT}_{224}$	4.4 5.9%	3.6 4.2%	3.6 4.4%	2.9 5.4%	5.8 9.0%	2.5 4.1%	8.8 10.3%	6.7 7.4%	5.7 6.7%	7.4 9.5%	7.1 8.2%	0.8 1.0%	14.7%(↓ 85.3%)
UTPTrack-S_{224}	72.3 98.2%	82.3 98.2%	79.3 98.2%	52.7 99.0%	63.7 98.8%	60.0 97.4%	85.2 99.1%	89.7 99.3%	84.4 98.5%	77.5 100.0%	87.1 100.5%	77.9 99.7%	**99.3**%(↓ 0.7%)

Table 17. Performance comparisons under different vision token compression configurations across **unified tracking** on RGB-D/T/E/Lang tracking benchmark. The vanilla number of visual tokens is 294. The first line of each method is the raw performance of benchmarks, and the second line is the proportion relative to the upper limit. The average performance listed is calculated across all 10 benchmarks.

Method	VOT-RGBD22			LasHeR		RGBT234		VisEvent		TNL2K		OTB99		Avg Perf.(%)
	EAO	Acc	Rob	AUC	P	MSR	MPR	AUC	P	AUC	P	AUC	P	
Upper Bound, 294 Tokens (100%)														
SUTrack_{224}	75.5 100%	82.5 100%	91.3 100%	59.9 100%	74.8 100%	70.0 100%	92.1 100%	63.3 100%	80.7 100%	67.8 100%	73.8 100%	68.4 100%	91.1 100%	100%
Retain 71.4% Tokens (≈ 218 tokens, ↓ 25.6%)														
SUTrack-CE_{224}	75.8 100.4%	82.8 100.4%	91.4 100.1%	59.9 100.0%	74.7 99.9%	70.0 100.0%	92.5 100.4%	61.8 97.6%	79.2 98.1%	66.0 97.3%	71.5 96.9%	70.5 103.0%	92.6 101.7%	<u>99.5%</u> (↓ 0.5%)
$\text{SUTrack-ToMe}_{224}$	76.0 100.7%	82.5 100.0%	91.9 100.7%	60.0 100.2%	74.6 99.7%	69.0 98.6%	92.0 99.9%	62.3 98.4%	79.6 98.6%	66.2 97.6%	71.8 97.3%	70.4 102.9%	91.3 100.3%	99.4%(↓ 0.6%)
$\text{SUTrack-EViT}_{224}$	75.7 100.3%	82.6 100.1%	91.4 100.1%	59.0 98.5%	73.2 97.9%	70.0 100.0%	92.0 99.9%	62.0 97.9%	79.1 98.0%	65.7 96.9%	71.1 96.4%	70.7 103.4%	91.4 100.4%	99.0%(↓ 1.0%)
$\text{SUTrack-DyViT}_{224}$	74.4 98.5%	81.5 98.8%	91.2 99.9%	57.6 96.2%	71.8 96.0%	68.7 98.1%	90.8 98.6%	61.3 96.8%	79.4 98.4%	64.0 94.4%	65.5 88.9%	71.5 104.5%	93.6 102.8%	96.5%(↓ 3.5%)
UTPTrack_{224}	76.5 101.3%	82.7 100.4%	92.1 100.9%	60.1 100.3%	74.9 100.1%	70.2 100.3%	93.1 101.1%	62.0 97.9%	79.2 98.1%	66.4 97.9%	72.0 97.6%	70.7 103.3%	93.1 102.2%	**99.8%**(↓ 0.2%)
Retain 52.0% Tokens (≈ 153 tokens, ↓ 48.0%)														
SUTrack-CE_{224}	76.1 100.8%	82.8 100.4%	91.8 100.5%	58.4 97.5%	72.7 97.2%	70.4 100.6%	93.0 101.0%	61.5 97.2%	78.6 97.4%	66.1 97.5%	71.7 97.2%	70.5 103.0%	92.6 101.7%	<u>99.2%</u>(↓ 0.8%)
$\text{SUTrack-ToMe}_{224}$	76.1 100.8%	82.7 100.2%	91.8 100.5%	59.2 98.8%	73.7 98.5%	69.4 99.1%	91.8 99.7%	61.7 97.5%	78.9 97.8%	65.9 97.1%	71.4 96.9%	70.5 103.1%	91.6 100.6%	99.0%(↓ 0.5%)
$\text{SUTrack-EViT}_{224}$	75.2 99.6%	82.1 99.5%	91.4 100.1%	57.8 96.5%	71.9 96.1%	69.9 99.9%	92.3 100.2%	62.3 98.4%	79.8 98.9%	65.1 96.0%	70.3 95.4%	70.5 103.1%	91.7 100.7%	98.5%(↓ 1.5%)
$\text{SUTrack-DyViT}_{224}$	63.0 83.4%	74.3 90.1%	84.8 92.9%	50.6 84.5%	63.4 84.8%	63.3 90.4%	86.3 93.7%	49.6 78.4%	67.2 83.3%	54.3 80.0%	54.4 73.7%	65.3 95.5%	86.0 94.4%	78.8%(↓ 21.2%)
UTPTrack_{224}	75.7 100.3%	82.8 100.4%	91.4 100.1%	59.6 99.5%	74.4 99.5%	70.7 101.0%	93.6 101.6%	61.7 97.5%	78.8 97.6%	65.9 97.2%	71.5 96.9%	70.1 102.5%	91.2 100.1%	**99.5%**(↓ 0.5%)
Retain 35.4% Tokens (≈ 104 tokens, ↓ 64.6%)														
SUTrack-CE_{224}	75.5 100.0%	82.2 99.6%	91.8 100.5%	57.8 96.5%	71.9 96.1%	69.3 99.0%	92.2 100.1%	61.4 97.0%	78.7 97.5%	65.4 96.3%	70.5 95.5%	71.7 104.8%	94.2 103.5%	<u>98.3%</u>(↓ 1.7%)
$\text{SUTrack-ToMe}_{224}$	70.9 93.9%	78.6 95.3%	90.2 98.8%	53.8 89.8%	66.6 89.0%	63.6 90.9%	87.9 95.4%	58.0 91.6%	77.4 95.9%	61.7 91.0%	63.6 86.3%	67.3 98.4%	87.0 95.6%	92.5%(↓ 7.5%)
$\text{SUTrack-EViT}_{224}$	75.0 99.3%	81.9 99.3%	91.1 99.8%	56.4 94.2%	70.3 94.0%	67.8 96.9%	90.2 97.9%	60.8 96.1%	78.4 97.1%	64.4 95.0%	69.3 94.0%	70.4 102.9%	91.5 100.5%	96.7%(↓ 3.3%)
$\text{SUTrack-DyViT}_{224}$	13.7 18.1%	49.7 60.2%	23.6 25.8%	10.3 17.2%	15.5 20.7%	17.9 25.6%	32.4 35.2%	7.3 11.5%	14.8 18.3%	15.9 23.4%	13.5 18.3%	13.6 19.9%	20.8 22.9%	14.7%(↓ 85.3%)
UTPTrack_{224}	76.1 100.8%	83.0 100.6%	91.5 100.2%	58.6 97.8%	72.9 97.5%	70.4 100.6%	92.9 100.9%	61.5 97.2%	79.1 98.0%	66.0 97.3%	71.4 96.9%	70.3 102.7%	91.9 100.9%	**99.3%**(↓ 0.7%)

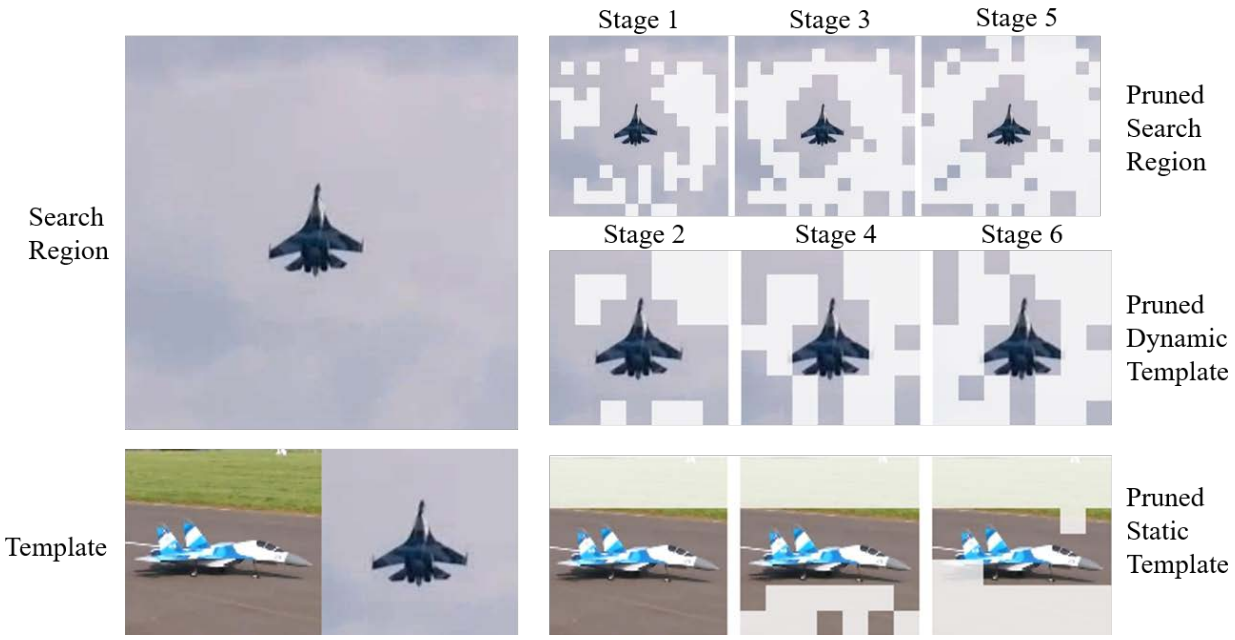

Figure 5. Visualization of the UTPTrack Pruning Process for RGB-based Tracking.

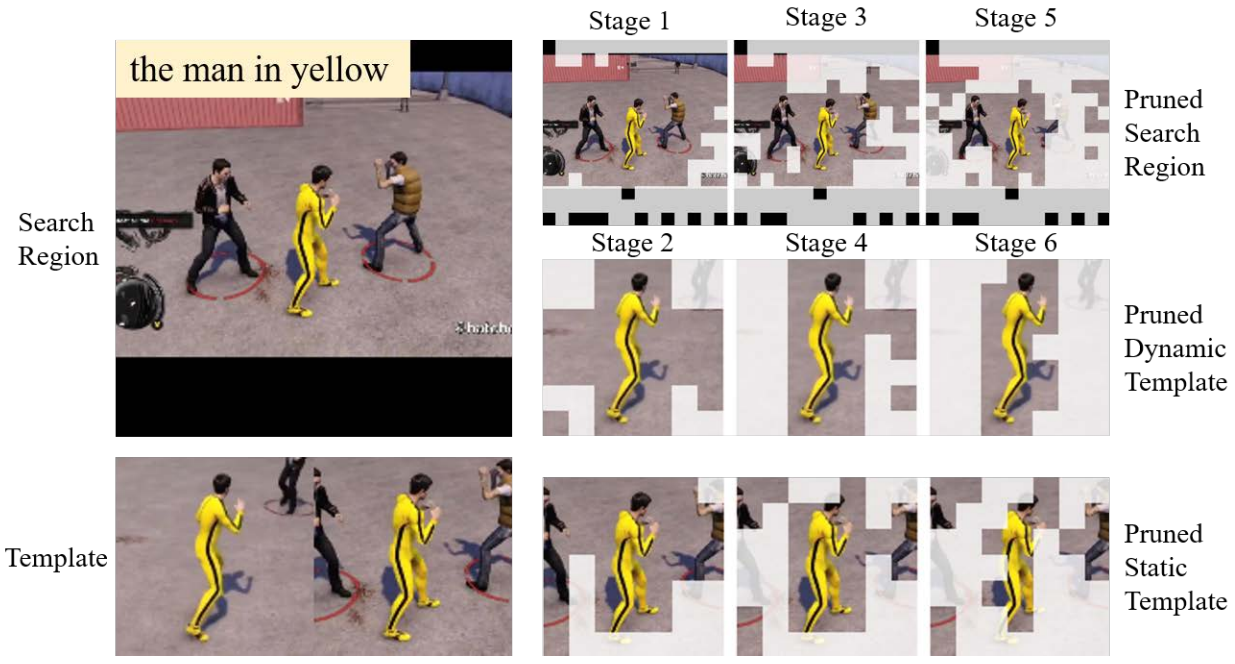

Figure 6. Visualization of the UTPTrack Pruning Process for RGB-Language Tracking.

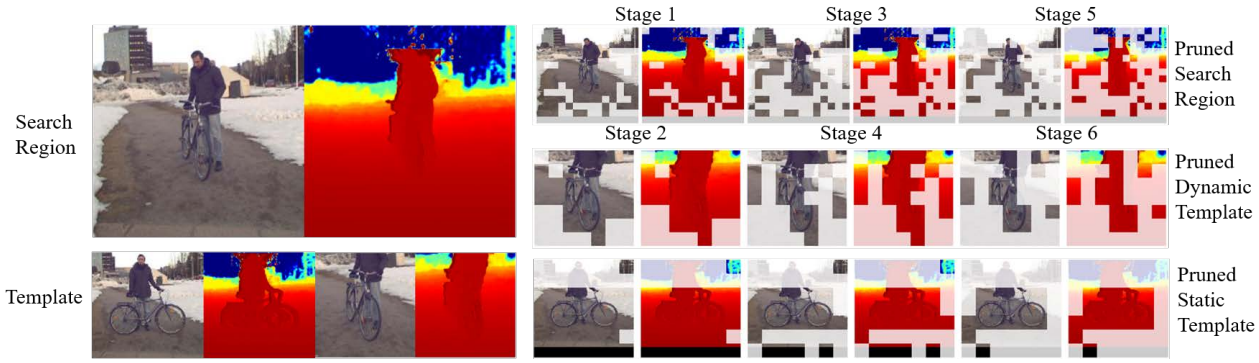

Figure 7. Visualization of the UTPTrack Pruning Process for RGB-Depth Tracking.

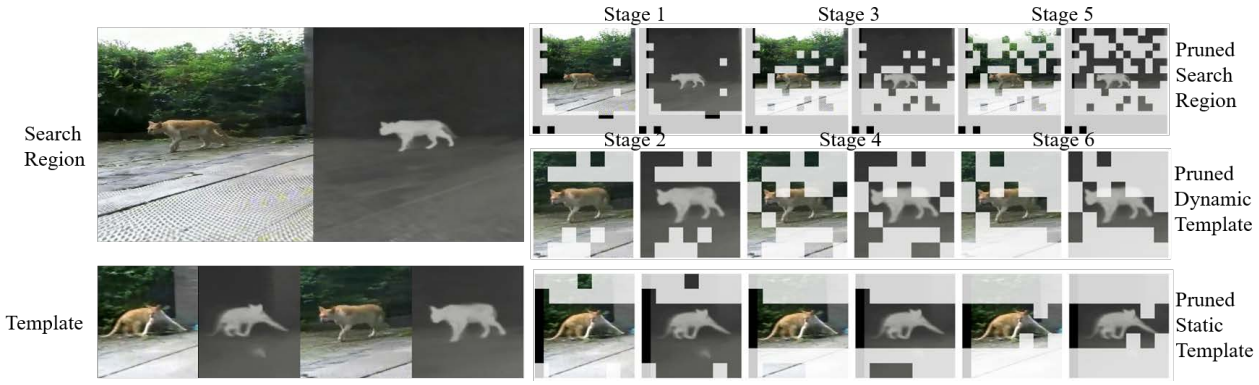

Figure 8. Visualization of the UTPTrack Pruning Process for RGB-Thermal Tracking.

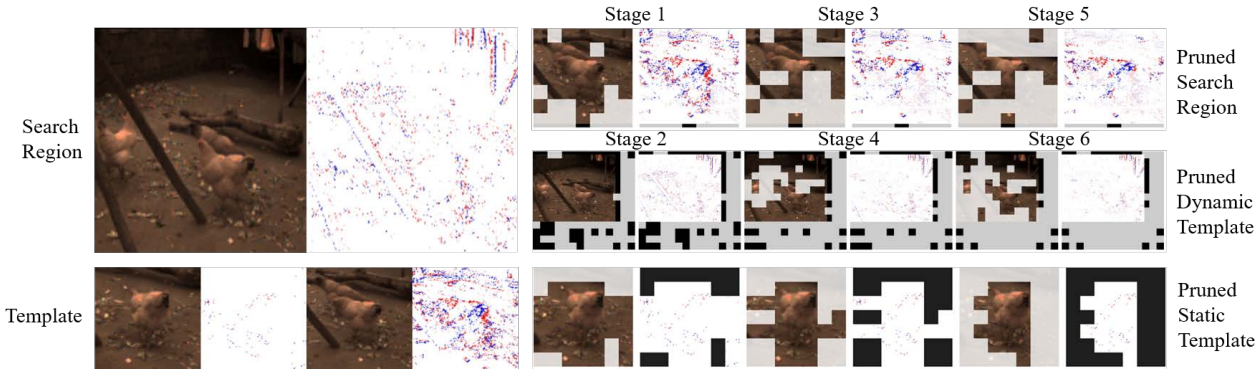

Figure 9. Visualization of the UTPTrack Pruning Process for RGB-Event Tracking.