AMBIGUOUS TOKEN REFINEMENT FOR TARGET REPRESENTATION IN TRANSFORMER TRACKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Transformer-based visual trackers have achieved remarkable performance through effective template–search interaction. However, current template token representation relies on bounding-box overlap as a foreground prior, implicitly assuming that all target-overlapping tokens are equally relevant to the target. This geometric assumption overlooks semantic ambiguity within the target region, where some overlapping tokens may exhibit background-like semantics and provide unreliable information for template modeling. In this paper, we propose **Ambiguous Token Refinement (ATR)** to address this limitation. ATR refines the geometry-derived foreground prior by introducing an additional ambiguous token type for tokens that overlap with the target box but are semantically closer to the surrounding background than the target object. By incorporating semantic relevance into token-type assignment, ATR reduces unreliable foreground cues while preserving informative target appearance. ATR is model-agnostic and can be seamlessly integrated into various tracking frameworks, including RGB, RGB-Thermal, RGB-Depth, RGB-Event, and RGB-Language tracking. Extensive experiments on multiple benchmarks demonstrate consistent improvements over strong baselines, validating ATR as a simple yet effective strategy for Transformer-based visual tracking.

1 INTRODUCTION

Transformer-based trackers have achieved strong performance by jointly modeling the template and search regions through powerful token interactions (Yan et al., 2021; Ye et al., 2022; Cui et al., 2022; Chen et al., 2023; Wei et al., 2023; Lin et al., 2024; 2026). Unlike video object segmentation (Cheng & Schwing, 2022; Wang et al., 2023a; Hong et al., 2023; Cheng et al., 2024; Liu et al., 2025), where pixel-level masks explicitly define object boundaries, tracking supervision is typically limited to a bounding box. As a result, a tracking template inevitably mixes target and background content, making precise target representation intrinsically difficult.

Most transformer trackers preserve background context because it can provide useful cues about the target and its local environment (Cui et al., 2022; Chen et al., 2023), as shown in Fig. 1(a-b). However, treating all template tokens without explicit target location and semantic distinction also allows irrelevant background content to participate in target modeling and template–search interaction. Recent work has therefore explored explicit foreground–background token representations. For example, LoRAT (Lin et al., 2024) projects the target bounding box onto the patch grid and assigns binary token types: tokens within or overlapping the box are treated as foreground, while the remaining tokens are treated as background, as illustrated in Fig. 1(c).

Although such binary token typing conflates *geometric overlap* with *semantic foregroundness*, a bounding box only specifies a coarse spatial extent of the target. Due to irregular object shapes, pose variations, and misalignment with the discrete patch grid, some tokens overlapping the box may contain substantial background content. Nevertheless, binary assignment labels all of them as foreground. Consequently, background-like tokens inside the coarse target region remain indistinguishable from genuine target tokens, limiting the accuracy of template representation.

We address this limitation with **Ambiguous Token Refinement (ATR)**. Our key observation is that box-based token assignment provides asymmetric information: tokens that do not overlap the target box form explicit non-target references, whereas box-overlapping tokens remain semantically uncertain. ATR exploits this asymmetry to refine the coarse foreground assignment. Specifically,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

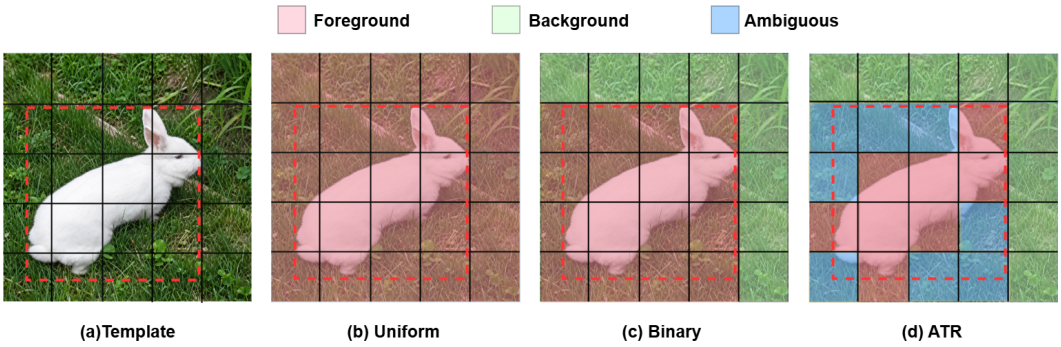


Figure 1: Comparison of template token representations. Conventional trackers (a-b) do not explicitly distinguish template token types, while recent methods (c) such as LoRAT (Lin et al., 2024) define foreground/background token types according to target bounding-box geometry. This binary assignment overlooks semantic ambiguity within the target region. Our method (d) refines the geometry-based prior by introducing an ambiguous type for tokens inside the target region that are semantically similar to the background.

we aggregate the explicit background tokens into a background prototype and compare each box-overlapping token with this reference using a similarity function. Tokens with higher similarity to the background reference are identified as ambiguous and refined accordingly, as illustrated in Fig. 1(d). In contrast to binary token typing that relies solely on spatial overlap, ATR incorporates appearance evidence to resolve the semantic ambiguity of box-overlapping tokens, yielding a fine-grained and more accurate target representation.

ATR is model-agnostic and is applied to the template image once before tracking, with no inference overhead. It can be directly integrated into different tracking architectures and naturally extends beyond standard RGB tracking to RGB+Thermal, RGB+Depth, RGB+Event, and RGB-Language settings. Extensive experiments across single-template, multi-frame, and multimodal tracking demonstrate consistent improvements over diverse baseline trackers, validating the generality and effectiveness of the proposed refinement.

Our main contributions can be summarized as follows:

1. We identify a limitation of binary token representation in transformer tracking: geometric box overlap does not necessarily imply semantic foregroundness, causing background-like content within the target box to be incorrectly represented as foreground.
2. We propose **Ambiguous Token Refinement (ATR)**, which leverages explicit background tokens as references to identify and refine semantically ambiguous box-overlapping tokens. ATR extends coarse geometry-based token typing with semantic-aware refinement for more accurate target representation.
3. ATR is lightweight, model-agnostic, and broadly applicable. Extensive experiments across single-template, multi-frame, and multimodal tracking demonstrate consistent improvements on diverse trackers without additional inference overhead.

2 Related Work

2.1 Transformer-based Tracking

Transformer-based visual trackers have recently achieved significant progress by leveraging the strong representation capability of Transformers (Liu et al., 2021; Dosovitskiy, 2020) and large-scale pretraining. Early Transformer trackers (Wang et al., 2021a) adopted a Siamese framework, where template and search features were extracted separately and fused through Transformer modules. Following this direction, TransT (Chen et al., 2021) and SwinTrack (Lin et al., 2022) further enhanced feature interaction with more powerful Transformer designs. However, these dual-branch frameworks perform feature extraction before template–search interaction, limiting joint representation learning.

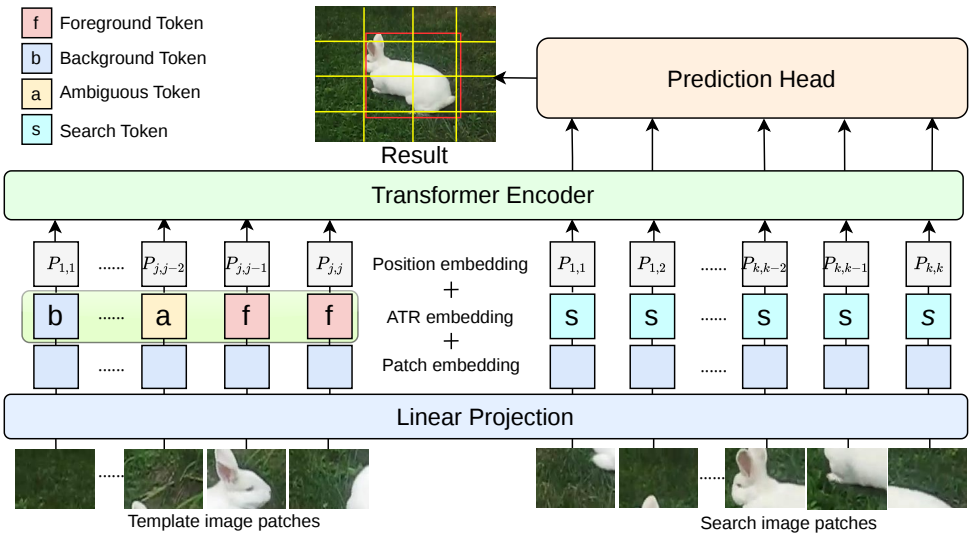


Figure 2: An overview of the proposed ATR method for Transformer-based tracking. Based on the target box and token semantics in the template image, we assign each token to foreground, background, or ambiguous type, providing richer information about the target and background. Our method is general, and it can be seamlessly integrated into various tracking frameworks.

Recent one-stream trackers, including MixFormer (Cui et al., 2022), OSTrack (Ye et al., 2022), and ODTrack (Zheng et al., 2024), jointly process template and search tokens within a unified Transformer backbone, achieving strong tracking performance. Meanwhile, LoRAT (Lin et al., 2024; 2026) explores parameter-efficient adaptation based on Low-Rank Adaptation (LoRA) to reduce the cost of fine-tuning large pretrained Transformers. In this work, we mainly focus on improving template token refinement for accurate target representation.

2.2 Template Modeling

Early patch-based trackers (Li et al., 2017; Mao et al., 2014; Li et al., 2015) represent targets using local patches and explicitly consider the reliability of different regions. For example, (Li et al., 2015) exploits reliable patches that can be consistently tracked throughout the tracking process, highlighting the importance of region-level target representation.

Modern deep learning-based trackers (Bertinetto et al., 2016; Cui et al., 2022; Gao et al., 2023; Cai et al., 2023; Zhang et al., 2025; Wang et al., 2019; Yan et al., 2021; Yang et al., 2023) typically crop a template region larger than the target bounding box to incorporate contextual information. While such context improves robustness to appearance variations, uniformly modeling template tokens may introduce irrelevant information into target representation. Recently, LoRAT (Lin et al., 2024) introduces foreground/background token-type embeddings based on target bounding-box overlap to provide explicit token semantics. However, this binary assignment still relies on geometric priors and cannot distinguish semantic ambiguity among target-related tokens. In this work, we further refine template token representation by introducing an additional ambiguous token type.

3 Method

3.1 Revisiting Template Token Representation

Given a template image $\mathbf{I}_T \in \mathbb{R}^{H_T \times W_T \times 3}$ from the initial frame and a search image $\mathbf{I}_S \in \mathbb{R}^{H_S \times W_S \times 3}$ from the video stream, Transformer-based trackers localize the target by jointly modeling template and search regions. Following the standard patch-based representation, the two images are divided into non-overlapping patches with patch size $P \times P$. This produces $N_T = H_T W_T / P^2$ template patches and $N_S = H_S W_S / P^2$ search patches (Ye et al., 2022), respectively. Each patch is then

projected into a d-dimensional visual token embedding:

$$\mathbf{X}_T \in \mathbb{R}^{N_T \times d}, \qquad \mathbf{X}_S \in \mathbb{R}^{N_S \times d}. \tag{1}$$

$\mathbf{X}_T$ and $\mathbf{X}_S$ denote template and search patch embeddings, respectively. Let $\mathbf{P}_T$ and $\mathbf{P}_S$ be position embeddings that preserve the spatial information of template and search tokens. They are added to the patch embeddings to denote the location of each patch as:

$$\mathbf{Z}_T = \mathbf{X}_T + \mathbf{P}_T, \qquad \mathbf{Z}_S = \mathbf{X}_S + \mathbf{P}_S, \tag{2}$$

The two token sequences are then concatenated and processed by a shared Transformer encoder:

$$[\mathbf{F}_T; \mathbf{F}_S] = \mathrm{Encoder}([\mathbf{Z}_T; \mathbf{Z}_S]), \tag{3}$$

where $\mathbf{F}_T$ and $\mathbf{F}_S$ denote the encoded template and search features. The search features $\mathbf{F}_S$ are used for target localization, while $\mathbf{F}_T$ provides target appearance information for template–search interaction. Finally, the $\mathbf{F}_S$ are reshaped to a 2D feature map according to their original spatial positions. This feature map is then fed into the prediction head to estimate the target bounding box.

Binary token representation. Since the template region contains both target appearance and surrounding context, template tokens may contribute differently to target representation. To distinguish their roles, recent methods introduce foreground/background token-type embeddings. Taking LoRAT (Lin et al., 2024) as an example, token types are assigned according to the spatial overlap between template patches and the target bounding box.

Let Ω denote the set of token positions whose patches overlap with the target bounding box. Following Eq. 2, the binary token representation is formulated as:

$$\mathbf{Z}_T^{(i,j)} = \begin{cases} \mathbf{X}_T^{(i,j)} + \mathbf{P}_T^{(i,j)} + \mathbf{E}_f^{(i,j)}, & (i,j) \in \Omega, \\ \mathbf{X}_T^{(i,j)} + \mathbf{P}_T^{(i,j)} + \mathbf{E}_b^{(i,j)}, & (i,j) \notin \Omega, \end{cases}$$
$$\mathbf{Z}_S^{(i,j)} = \mathbf{X}_S^{(i,j)} + \mathbf{P}_S^{(i,j)} + \mathbf{E}_s^{(i,j)}, \tag{4}$$

where $\mathbf{E}_f^{(i,j)}, \mathbf{E}_b^{(i,j)} \in \mathbb{R}^d$ are foreground and background token-type embeddings, respectively. $\mathbf{E}_s^{(i,j)} \in \mathbb{R}^d$ denotes the token-type embedding for search tokens.

Although this binary representation provides an effective spatial prior, it assumes that all tokens overlapping with the target bounding box are equally related to the target. However, bounding-box overlap only indicates spatial location and does not necessarily reflect semantic relevance. Some target-overlapping tokens may contain background-like patterns and provide unreliable target cues when treated as foreground.

3.2 Ambiguous Token Refinement

The above observation motivates us to move beyond binary token assignment based solely on bounding-box overlap. We propose **Ambiguous Token Refinement (ATR)**, which refines the geometry-derived foreground prior by modeling semantic ambiguity. Instead of treating all target-overlapping tokens as foreground, ATR identifies background-like tokens among them and assigns an additional ambiguous type.

Ambiguous token recognition. ATR estimates token ambiguity by measuring the similarity between each template token and the explicit background regions outside the target bounding box. Based on an ambiguity score a_{ij}, each template token is assigned to one of three semantic types: foreground, ambiguous, or background. Then the template token in Eq. 4 is rewritten as:

$$\mathbf{Z}_T^{(i,j)} = \begin{cases} \mathbf{X}_T^{(i,j)} + \mathbf{P}_T^{(i,j)} + \mathbf{E}_f^{(i,j)}, & (i,j) \in \Omega,\ a_{ij} \le \tau, \\ \mathbf{X}_T^{(i,j)} + \mathbf{P}_T^{(i,j)} + \mathbf{E}_a^{(i,j)}, & (i,j) \in \Omega,\ a_{ij} > \tau, \\ \mathbf{X}_T^{(i,j)} + \mathbf{P}_T^{(i,j)} + \mathbf{E}_b^{(i,j)}, & (i,j) \notin \Omega, \end{cases} \tag{5}$$

$\mathbf{E}_f^{(i,j)}, \mathbf{E}_a^{(i,j)}, \mathbf{E}_b^{(i,j)} \in \mathbb{R}^d$ denote the foreground, ambiguous, and background token-type embeddings, respectively. $\mathbf{E}_b^{(i,j)}$ is inherited from the original binary representation. τ denotes the ambiguity threshold. Notice that $\tau = 1$ indicates binary token representation. For target-overlapping tokens, $a_{ij} \leq \tau$ retains the foreground type, while $a_{ij} > \tau$ assigns the ambiguous type. The ambiguity score a_{ij} is computed by the background-guided estimation above.

Ambiguity probability estimation. ATR uses tokens outside the target bounding box as background references. Specifically, template tokens whose patches do not overlap with the target box are used to construct the background prototype $\mathbf{X}_B$, providing contextual background information for ambiguity estimation. Let N_b denote the number of non-target tokens, $\mathbf{X}_B$ is formulated as:

$$\mathbf{X}_B = \frac{1}{N_b} \sum_{(i,j) \notin \Omega} \mathbf{X}_T^{(i,j)}, \tag{6}$$

For each template token, the ambiguity score a_{ij} is computed by a cosine similarity function as:

$$a_{ij} = \frac{1}{2} \left(1 + \frac{(\mathbf{X}_T^{(i,j)})^\top \mathbf{X}_B}{\|\mathbf{X}_T^{(i,j)}\|_2 \|\mathbf{X}_B\|_2 + \epsilon} \right), \tag{7}$$

where ϵ is a small constant for numerical stability. The score $a_{ij} \in [0, 1]$ measures the background similarity of token $\mathbf{X}_T^{(i,j)}$.

ATR only refines token-type assignments according to the estimated ambiguity scores, while leaving the visual token embeddings and all other tracking modules unchanged. It requires no additional training objective or architectural modification, making it readily applicable to existing Transformer-based tracking frameworks. Moreover, ATR is applied to the template image once before tracking, and thus there is no additional inference computational cost introduced.

4 Experiments

4.1 Experimental Setup

Implementation Details. We evaluate Ambiguous Token Refinement (ATR) as a plug-in module on three types of trackers covering visible and multi-modal tracking settings: (1) single-template tracker LoRAT (Lin et al., 2024), which adopts binary foreground/background token-type embeddings; (2) multi-frame trackers, including DTPTrack (Huang et al., 2026) and UTPTrack (Wu et al., 2026), which augment existing trackers with plug-in modules, allowing us to evaluate whether ATR provides additional benefits alongside these modules; (3) multi-modal tracker SUTrack (Chen et al., 2025), which unifies RGB, RGB-Thermal, RGB-Depth, RGB-Event, and RGB-Language tracking in a jointly trained model, aiming to assess the applicability of ATR across different input modalities. Specifically, for each baseline tracker, we keep its original architecture, training strategy, and evaluation protocol unchanged, while only incorporating the proposed token representation module into the template encoding stage. The threshold for ambiguous-token assignment is fixed to $\tau = 0.6$ for all our experiments. All models are trained on 8 NVIDIA RTX 4090D GPUs and evaluated on one GPU.

4.2 Main Results

Table 1 summarizes the comparisons between the baselines and their ATR-enhanced versions. It can be seen that ATR provides consistent performance gains across different baseline trackers, including RGB and multi-modal tracking settings, demonstrating our method's strong generalization to various tracking tasks. We discuss the results for each tracking setting below.

Single-template Tracking. As a fundamental and widely studied tracking paradigm, single-template tracking has been extensively evaluated on a diverse set of established benchmarks. Accordingly, we conduct comprehensive comparisons on seven RGB-based benchmarks, including five challenging large-scale benchmarks and two widely used smaller-scale benchmarks. Table 2 reports the results on LaSOT (Fan et al., 2019), LaSOT$_{\text{ext}}$ (Fan et al., 2021), GOT-10k (Huang et al., 2019),

Table 1: Generalization of ATR across different tracking settings. Bold indicates improvements over the corresponding baseline. "–" denotes unreported results.

Single-template tracking	LaSOT			$\text{LaSOT}_{\text{ext}}$			GOT-10k			VastTrack	
	AUC	P_{Norm}	P	AUC	P_{Norm}	P	AO	$\text{SR}_{0.5}$	$\text{SR}_{0.75}$	AUC	P
LoRAT-B224 (Lin et al., 2024)	71.7	80.9	77.3	50.3	61.6	57.1	72.1	81.8	70.7	38.7	37.8
LoRAT-B224 + ATR	**72.7**	**82.3**	**78.3**	**50.9**	**62.3**	**57.9**	**76.1**	**83.9**	**72.4**	**39.4**	**38.5**
Multi-frame tracking	LaSOT			GOT-10k			TrackingNet		VastTrack	
	AUC	P_{Norm}	P	AO	$\text{SR}_{0.5}$	$\text{SR}_{0.75}$	AUC	P_{Norm}	AUC	P
DTPTrack-B224 (Huang et al., 2026)	74.3	83.6	80.7	76.5	87.6	74.9	85.6	90.1	40.7	40.2
DTPTrack-B224 + ATR	**74.8**	**84.4**	**81.2**	**76.8**	**87.9**	**75.3**	**85.8**	**90.2**	**41.1**	**40.8**
UTPTrack-B256 (Wu et al., 2026)	67.4	76.3	72.5	75.2	85.3	73.9	84.0	88.8	–	–
UTPTrack-B256 + ATR	**67.8**	**76.6**	**72.9**	**75.8**	**85.7**	**74.3**	**84.1**	**88.9**	–	–
Multimodal tracking	VOT-RGBD22			LasHeR		VisEvent		TNL2K	
	EAO	Acc.	Rob.	AUC	P	AUC	P	AUC	P
SUTrack-B224 (Chen et al., 2025)	76.5	82.8	91.8	59.9	74.5	62.7	79.9	65.0	67.9
SUTrack-B224 + ATR	**76.7**	**82.9**	**92.0**	**60.5**	**74.9**	**63.3**	**80.5**	**65.8**	**68.6**

Table 2: State-of-the-art comparisons of single-template tracking.

Single-template Tracking	LaSOT			LaSOText			GOT-10k			TNL2K		VastTrack	
	AUC	P_{Norm}	P	AUC	P_{Norm}	P	AO	$\text{SR}_{0.5}$	$\text{SR}_{0.75}$	AUC	P	AUC	P
SeqTrack-B256 (Chen et al., 2023)	69.9	79.7	76.3	49.5	60.8	56.3	74.7	84.7	71.8	56.4	–	–	–
ARTrack-B256 (Wei et al., 2023)	70.4	79.5	76.6	46.4	56.5	52.3	73.5	82.2	70.9	**59.8**	–	–	–
ROMTrack-B256 (Cai et al., 2023)	69.3	78.8	75.6	48.9	59.3	55.0	72.9	82.9	70.2	–	–	–	–
AQATrack-B256 (Xie et al., 2024)	71.4	81.9	**78.6**	**51.2**	62.2	**58.9**	73.8	83.2	72.1	57.8	59.4	–	–
ARTrackV2-B256 (Bai et al., 2024)	71.6	80.2	77.2	50.8	61.9	57.7	75.9	85.4	72.7	59.2	–	–	–
LMTrack-B256 (Xu et al., 2025)	69.8	79.2	76.3	49.0	59.6	55.8	**76.3**	**87.1**	**73.9**	–	–	–	–
MIMTrack-B256 (Wang et al., 2025a)	69.1	78.8	75.7	47.7	57.0	54.3	72.6	83.2	69.3	–	–	–	–
LoRAT-B224 (Lin et al., 2024)	71.7	80.9	77.3	50.3	61.6	57.1	72.1	81.8	70.7	58.8	61.3	38.7	37.8
LoRAT-B224 + ATR	**72.7**	**82.3**	78.3	50.9	**62.3**	57.9	76.1	83.9	72.4	59.2	**61.8**	**39.4**	**38.5**
ROMTrack-384 (Cai et al., 2023)	71.4	81.4	78.2	51.3	62.4	58.6	74.2	84.3	72.4	–	–	37.0	36.1
SeqTrack-L384 (Chen et al., 2023)	72.5	81.5	79.3	50.7	61.6	57.5	74.8	81.9	72.2	57.8	–	39.6	40.2
ARTrack-L384 (Wei et al., 2023)	73.1	82.2	80.3	52.8	62.9	59.7	78.5	87.4	77.8	60.3	–	35.6	32.4
ARTrackV2-L384 (Bai et al., 2024)	73.6	82.8	81.1	53.4	63.7	60.2	**79.5**	**87.8**	**79.6**	61.6	–	–	–
ODTrack-L384 (Zheng et al., 2024)	74.0	84.2	82.3	53.9	65.4	61.7	78.2	87.2	77.3	61.7	–	–	–
MIMTrack-L384 (Wang et al., 2025a)	71.6	82.0	79.7	49.6	60.3	57.7	75.3	85.8	72.5	–	–	–	–
TemTrack-384 (Xie et al., 2025)	73.1	83.0	80.7	53.4	64.8	61.0	76.1	84.9	74.4	–	–	–	–
MambaLCT-L384 (Li et al., 2025)	73.6	84.1	81.6	53.3	64.8	61.4	76.2	86.7	74.3	58.5	–	–	–
LoRAT-L378 (Lin et al., 2024)	75.1	84.1	82.0	56.6	69.0	65.1	77.5	86.2	78.1	62.3	67.0	43.9	45.8
LoRAT-L378 + ATR	**75.7**	**84.7**	**82.8**	**57.4**	**69.6**	**65.8**	78.0	86.7	78.5	**62.7**	**67.6**	**44.4**	**46.3**

Table 3: Performance comparison on UAV123 tracking benchmark.

Dataset	CTTrack (Song et al., 2023)	GRM (Gao et al., 2023)	MixViT (Yu et al., 2023)	SeqTrack (Chen et al., 2023)	ARTrackV2 (Bai et al., 2024)	MambaLCT (Li et al., 2025)	LoRAT (Lin et al., 2024)	LoRAT+ATR
UAV123	71.3	70.2	68.7	68.5	69.9	70.1	71.9	**72.5**

Table 4: Performance comparison on OTB100 tracking benchmark.

Dataset	TransT (Chen et al., 2021)	STARK (Yan et al., 2021)	AiATrack (Gao et al., 2022)	SBT (Xie et al., 2022)	OSTrack (Ye et al., 2022)	ADSTrack (Wang et al., 2025b)	LoRAT (Lin et al., 2024)	LoRAT+ATR
OTB100	69.4	68.0	69.6	70.9	70.9	68.4	71.5	**71.9**

TNL2K (Wang et al., 2021b), and VastTrack (Peng et al., 2024). Incorporating the proposed ATR consistently improves both LoRAT-B224 and LoRAT-L378 across all reported metrics, demonstrating its effectiveness across different model capacities.

Table 5: State-of-the-art comparisons of multi-frame tracking.

Multi-frame tracking	LaSOT			GOT-10k			TrackingNet		VastTrack	
	AUC	P_{Norm}	P	AO	$SR_{0.5}$	$SR_{0.75}$	AUC	P_{Norm}	AUC	P
HIPTrack (Cai et al., 2024)	72.7	82.9	79.5	77.4	**88.0**	74.5	84.5	89.1	–	–
ARTrackV2-B384 (Bai et al., 2024)	73.0	82.0	79.6	**77.5**	86.0	**75.5**	85.7	89.8	–	–
AQATrack-B-384 (Xie et al., 2024)	72.7	82.9	80.2	76.0	85.2	74.9	84.8	89.3	–	–
ODTrack-B384 (Zheng et al., 2024)	73.2	83.2	80.6	77.0	87.9	75.1	85.1	**90.1**	–	–
LoRAT-B378 (Lin et al., 2024)	72.9	81.9	79.1	73.7	82.6	72.9	84.2	88.4	38.7	37.8
TemTrack-B256 (Xie et al., 2025)	72.0	82.1	79.1	74.9	84.8	71.7	84.3	88.8	–	–
LoRATv2-B224 (Lin et al., 2026)	72.0	81.3	77.9	74.7	84.7	72.7	84.1	–	39.1	38.7
UTPTrack-B256 (Wu et al., 2026)	67.4	76.3	72.5	75.2	85.3	73.9	84.0	88.8	–	–
DTPTrack-B224 (Huang et al., 2026)	74.3	83.6	80.7	76.5	87.6	74.9	85.6	**90.1**	40.7	40.2
UTPTrack-B256 + ATR	67.8	76.6	72.9	75.8	85.7	74.3	84.1	88.9	–	–
DTPTrack-B224 + ATR	**74.8**	**84.4**	**81.2**	76.8	87.9	75.3	**85.8**	**90.1**	**41.1**	**40.8**
OSTrack-L384 (Ye et al., 2022)	71.1	81.1	77.6	73.7	83.2	70.8	83.9	88.5	33.6	31.5
ARTrack-L384 (Wei et al., 2023)	73.1	82.2	80.3	78.5	87.4	77.8	85.6	89.6	35.6	32.4
ARTrackV2-L384 (Bai et al., 2024)	73.6	82.8	81.1	79.5	87.8	79.6	86.1	90.4	–	–
LoRAT-L378 (Lin et al., 2024)	75.1	84.1	82.0	77.5	86.2	78.1	85.6	89.7	43.9	45.8
LoRAT-g378 (Lin et al., 2024)	76.2	85.3	83.5	78.9	87.8	80.7	86.0	90.2	46.0	48.8
SPMTrack-L378 (Cai et al., 2025)	76.8	85.9	84.0	80.0	89.4	79.9	**86.9**	**91.0**	–	–
LoRATv2-L378 (Lin et al., 2026)	76.1	85.1	83.1	78.2	86.8	79.1	85.7	–	44.2	46.7
UTPTrack-B384 (Wu et al., 2026)	70.7	80.6	77.2	76.5	86.9	74.5	84.0	88.7	–	–
DTPTrack-L378 (Huang et al., 2026)	77.2	86.3	84.2	80.3	89.9	81.9	86.5	90.8	46.6	49.1
UTPTrack-B384 + ATR	71.9	81.9	78.4	77.2	87.7	75.1	84.3	89.0	–	–
DTPTrack-L378 + ATR	**77.6**	**86.6**	**84.7**	**80.5**	**90.0**	**82.2**	86.6	90.9	**47.1**	**49.5**

We further evaluate LoRAT-B224 + ATR on two widely used smaller-scale benchmarks, UAV123 and OTB100. As shown in Tables 3 and 4, LoRAT-B224 + ATR achieves AUC scores of 72.5% and 71.9%, respectively.These results show that the benefit of ATR is also consistently observed on smaller tracking benchmarks.

Multi-frame Tracking. In multi-frame trackers, we apply ATR only to the first template frame, refining its token types into foreground, ambiguous, and background, while leaving the token-type assignments of all remaining frames unchanged. As shown in Table 5, DTPTrack-L378 + ATR achieves strong performance on LaSOT, GOT-10k, TrackingNet, and VastTrack. ATR also improves both the B256 and B384 variants of UTPTrack across different benchmarks. These results show that refining the initial template representation alone can benefit multi-frame tracking, complementing the temporal information already exploited by the baseline trackers.

Multi-modal Tracking. To further validate the applicability of ATR beyond RGB-only tracking, we evaluate it within the unified SUTrack framework on VOT-RGBD22 (Kristan et al., 2022), LasHeR (Li et al., 2021), VisEvent (Wang et al., 2023b), and TNL2K (Wang et al., 2021b), covering RGB-D, RGB-T, RGB-Event, and RGB-Language inputs, respectively. As shown in Table 6, ATR improves both SUTrack variants on all four benchmarks, with no degradation in any reported metric. Both ATR-enhanced variants achieve the best or joint-best results within their respective comparison groups. In particular, SUTrack-L378 + ATR improves AUC by 0.5 percentage points on LasHeR, VisEvent, and TNL2K. These gains suggest that refining ambiguous template tokens remains beneficial even when complementary modalities are available, supporting the effectiveness of ATR within a unified tracking framework.

4.3 Ablation Study

Effectiveness of Ambiguous Token Refinement. To validate the benefit of distinguishing semantically ambiguous box-overlapping tokens, we compare ATR with the binary token-type embedding (BTE) strategy in LoRAT-L224, which assigns all such tokens to the foreground type based solely on geometry. As shown in Table 7, replacing BTE in LoRAT-L224 with our ATR improves AUC

Table 6: State-of-the-art comparisons of multi-modal tracking.

Multi-modal tracking	VOT-RGBD22			LasHeR		VisEvent		TNL2K	
	EAO	Acc	Rob	AUC	P	AUC	P	AUC	P
OneTracker (Hong et al., 2024)	72.7	81.9	87.2	53.8	67.2	60.8	76.7	58.0	59.1
SeqTrackv2-B256 (Chen et al., 2024)	74.4	81.5	91.0	55.8	70.4	61.2	78.2	57.5	59.7
UniSOT-B256 (Ma et al., 2025)	73.2	81.8	88.8	54.0	67.4	59.2	76.2	58.0	60.3
UETrack-B224 (Kang et al., 2026)	68.3	80.8	-	55.5	69.1	60.7	78.0	53.9	54.4
SUTrack-B224 (Chen et al., 2025)	76.5	82.8	91.8	59.9	74.5	62.7	79.9	65.0	67.9
SUTrack-B224 + ATR	**76.7**	**82.9**	**92.0**	**60.5**	**74.9**	**63.3**	**80.5**	**65.8**	**68.6**
OneTracker (Hong et al., 2024)	72.7	81.9	87.2	53.8	67.2	60.8	76.7	58.0	59.1
SeqTrackv2-L384 (Chen et al., 2024)	74.8	82.6	91.0	61.0	76.7	63.4	80.0	62.4	66.1
UniSOT-L256 (Ma et al., 2025)	73.8	82.1	89.1	54.8	68.4	62.0	79.0	56.4	57.5
SUTrack-L378 (Chen et al., 2025)	76.6	83.5	92.2	61.9	76.9	63.8	80.5	67.9	72.1
SUTrack-L378 + ATR	**76.8**	**83.6**	**92.2**	**62.4**	**77.4**	**64.3**	**80.7**	**68.4**	**72.7**

Table 7: Effectiveness of ATR on LoRAT-L224 and UTPTrack-B384. BTE denotes the binary token embedding strategy.

Tracker	BTE	ATR	LaSOT AUC	$\text{LaSOT}_{\text{ext}}$ AUC
LoRAT-L224	×	×	74.0	52.4
	✓	×	74.2	52.8
	×	✓	**74.8**	**53.5**
UTPTrack-B384	×	×	70.7	49.4
	✓	×	71.0	50.0
	×	✓	**71.9**	**51.9**

Table 8: Effect of the ambiguity threshold τ in ATR using LoRAT-B224 as the baseline tracker.

τ	LaSOT AUC	$\text{LaSOT}_{\text{ext}}$ AUC	TNL2K AUC
0.3	71.8	50.3	58.8
0.4	71.9	50.3	58.8
0.5	72.1	50.6	59.0
0.6	**72.7**	50.7	**59.2**
0.7	72.4	**50.9**	58.9
0.8	72.0	50.3	58.9
1.0	71.7	50.3	58.8

by 0.6 and 0.7 percentage points on LaSOT and $\text{LaSOT}_{\text{ext}}$, respectively. For UTPTrack-B384, we introduce foreground/background token-type embeddings as a BTE variant and compare it with the proposed ATR strategy. Compared with the original UTPTrack-B384, BTE improves AUC by 0.3 and 0.6 percentage points on LaSOT and $\text{LaSOT}_{\text{ext}}$, respectively, whereas ATR achieves larger gains of 1.2 and 2.5 percentage points.

The consistent gains over BTE on both trackers support our motivation that geometric overlap alone is insufficient for template token typing. Refining this geometric prior using background similarity provides a more effective template representation by explicitly distinguishing background-like tokens within the coarse target region.

Effect of the Ambiguity Threshold τ. Using LoRAT-B224 as the baseline tracker, we investigate the influence of the ambiguity threshold τ, which controls the assignment of in-box tokens to foreground or ambiguous types. When $\tau = 1$, ATR reduces to the binary foreground/background assignment used in LoRAT, with all in-box tokens assigned to the foreground type. We evaluate $\tau \in \{0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 1.0\}$ on LaSOT, $\text{LaSOT}_{\text{ext}}$, and TNL2K with all other settings unchanged. As shown in Table 8, the improvements over $\tau = 1$ support the benefit of distinguishing background-like in-box tokens through an explicit ambiguous type.

4.4 Qualitative Results

Figure 3 compares LoRAT (Lin et al., 2024) and SUTrack (Chen et al., 2025) with their ATR-enhanced counterparts on representative RGB, RGB-Language, and RGB-D sequences. Green, red, and blue boxes denote the ground truth, ATR-enhanced predictions, and baseline predictions, respectively. As shown in Figure 3, LoRAT+ATR produces more accurate localization results than LoRAT by refining template token representations in RGB sequences. For RGB-Language and RGB-D tracking,

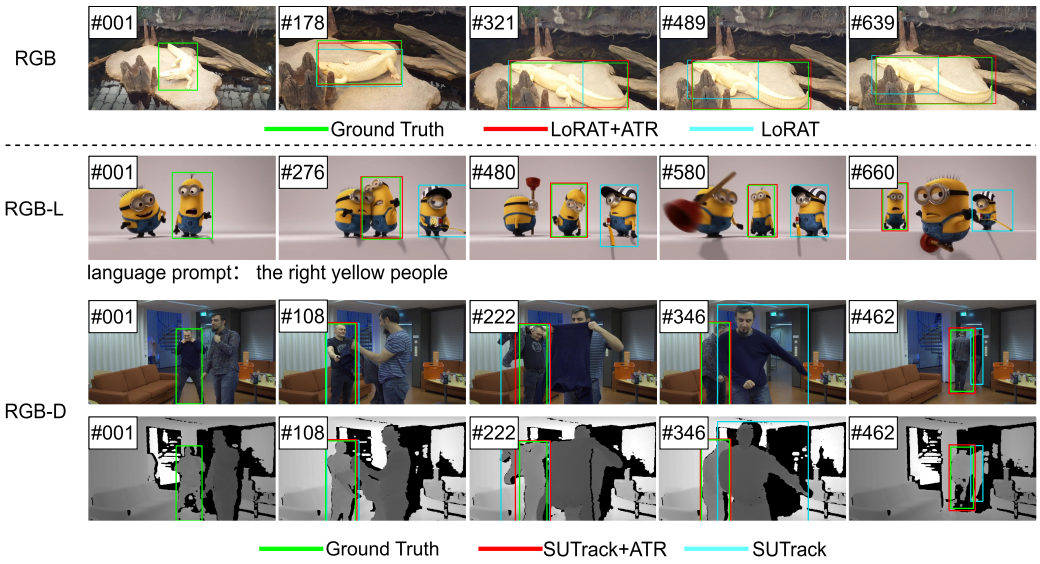


Figure 3: Qualitative comparison of LoRAT (Lin et al., 2024) and SUTrack (Chen et al., 2025) with their ATR-enhanced counterparts on challenging RGB, RGB-Language, and RGB-D sequences. ATR-enhanced trackers maintain more accurate target localization in the illustrated challenging scenarios.

SUTrack+ATR maintains more reliable target localization under complex scenarios. These results demonstrate that ATR reduces semantic interference in template modeling and improves tracking robustness across different input modalities.

5 Conclusion

In this paper, we introduced **Ambiguous Token Refinement (ATR)** to address the semantic ambiguity overlooked by geometry-based binary token assignment in Transformer tracking. ATR constructs a background prototype from explicit background tokens and assigns an additional ambiguous type to target-box-overlapping tokens with high background similarity. In this way, ATR preserves the spatial prior provided by the target bounding box while refining it with appearance information to better distinguish background-like content within the coarse target region. ATR is lightweight and can be integrated into template token encoding without modifying the backbone, prediction head, or training objective. Extensive experiments across single-template, multi-frame, and multimodal tracking demonstrate that ATR consistently improves strong baselines, supporting the effectiveness and generality of ambiguity-aware template token refinement.

AI use statement

AI assistance was used for grammar checking and sentence refinement to improve the clarity and readability of the manuscript. All scientific content, methodology, experiments, results, and conclusions were developed and verified by the authors. The authors reviewed the final manuscript and take full responsibility for all content presented in this work.

REFERENCES

Yifan Bai, Zeyang Zhao, Yihong Gong, and Xing Wei. Artrackv2: Prompting autoregressive tracker where to look and how to describe. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19048–19057, 2024.

Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pp. 850–865. Springer, 2016.

Wenrui Cai, Qingjie Liu, and Yunhong Wang. Hiptrack: Visual tracking with historical prompts. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19258–19267. IEEE, 2024.

Wenrui Cai, Qingjie Liu, and Yunhong Wang. Spmtrack: Spatio-temporal parameter-efficient fine-tuning with mixture of experts for scalable visual tracking. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16871–16881. IEEE, 2025.

Yidong Cai, Jie Liu, Jie Tang, and Gangshan Wu. Robust object modeling for visual tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9589–9600, 2023.

Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8126–8135, 2021.

Xin Chen, Houwen Peng, Dong Wang, Huchuan Lu, and Han Hu. Seqtrack: Sequence to sequence learning for visual object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14572–14581, 2023.

Xin Chen, Ben Kang, Jiawen Zhu, Dong Wang, Houwen Peng, and Huchuan Lu. Unified sequence-to-sequence learning for single- and multi-modal visual object tracking, 2024. URL https://arxiv.org/abs/2304.14394.

Xin Chen, Ben Kang, Wanting Geng, Jiawen Zhu, Yi Liu, Dong Wang, and Huchuan Lu. Sutrack: Towards simple and unified single object tracking. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 2239–2247, 2025.

Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European conference on computer vision*, pp. 640–658. Springer, 2022.

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3151–3161. IEEE, 2024.

Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13608–13618, 2022.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5374–5383, 2019.

Heng Fan, Hexin Bai, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Harshit, Mingzhen Huang, Juehuan Liu, et al. Lasot: A high-quality large-scale single object tracking benchmark. *International Journal of Computer Vision*, 129:439–461, 2021.

Shenyuan Gao, Chunluan Zhou, Chao Ma, Xinggang Wang, and Junsong Yuan. Aiatrack: Attention in attention for transformer visual tracking. In *European conference on computer vision*, pp. 146–164. Springer, 2022.

Shenyuan Gao, Chunluan Zhou, and Jun Zhang. Generalized relation modeling for transformer tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18686–18695, 2023.

Lingyi Hong, Wenchao Chen, Zhongying Liu, Wei Zhang, Pinxue Guo, Zhaoyu Chen, and Wenqiang Zhang. Lvos: A benchmark for long-term video object segmentation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13434–13446. IEEE, 2023.

Lingyi Hong, Shilin Yan, Renrui Zhang, Wanyun Li, Xinyu Zhou, Pinxue Guo, Kaixun Jiang, Yiting Chen, Jinglun Li, Zhaoyu Chen, and Wenqiang Zhang. Onetracker: Unifying visual object tracking with foundation models and efficient tuning, 2024. URL https://arxiv.org/abs/2403.09634.

Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43: 1562–1577, 2019.

Yuqing Huang, Liting Lin, Weijun Zhuang, Zhenyu He, and Xin Li. Drift-resilient temporal priors for visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6847–6856, June 2026.

Ben Kang, Jie Zhao, Xin Chen, Wanting Geng, Bin Zhang, Lu Zhang, Dong Wang, and Huchuan Lu. Uetrack: A unified and efficient framework for single object tracking. In *CVPR*, 2026.

Matej Kristan, Aleš Leonardis, Jiří Matas, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kämäräinen, Hyung Jin Chang, Martin Danelljan, Luka Čehovin Zajc, Alan Lukežič, et al. The tenth visual object tracking vot2022 challenge results. In *European Conference on Computer Vision*, pp. 431–460. Springer, 2022.

Chenglong Li, Liang Lin, Wangmeng Zuo, and Jin Tang. Learning patch-based dynamic graph for visual tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Chenglong Li, Wanlin Xue, Yaqing Jia, Zhichen Qu, Bin Luo, Jin Tang, and Dengdi Sun. Lasher: A large-scale high-diversity benchmark for rgbt tracking. *IEEE Transactions on Image Processing*, 31:392–404, 2021.

Xiaohai Li, Bineng Zhong, Qihua Liang, Guorong Li, Zhiyi Mo, and Shuxiang Song. Mambalct: Boosting tracking via long-term context state space model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4986–4994, 2025.

Yang Li, Jianke Zhu, and Steven CH Hoi. Reliable patch trackers: Robust visual tracking by exploiting reliable patches. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 353–361, 2015.

Liting Lin, Heng Fan, Zhipeng Zhang, Yong Xu, and Haibin Ling. Swintrack: A simple and strong baseline for transformer tracking. *Advances in Neural Information Processing Systems*, 35: 16743–16754, 2022.

Liting Lin, Heng Fan, Zhipeng Zhang, Yaowei Wang, Yong Xu, and Haibin Ling. Tracking meets lora: Faster training, larger model, stronger performance. In *European Conference on Computer Vision*, pp. 300–318. Springer, 2024.

Liting Lin, Heng Fan, Zhipeng Zhang, Yuqing Huang, Yaowei Wang, Yong Xu, and Haibin Ling. Loratv2: Enabling low-cost temporal modeling in one-stream trackers. *Advances in Neural Information Processing Systems*, 38:119714–119732, 2026.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Qin Liu, Jianfeng Wang, Zhengyuan Yang, Linjie Li, Kevin Lin, Marc Niethammer, and Lijuan Wang. Livos: Light video object segmentation with gated linear matching. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8668–8678. IEEE, 2025.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Yinchao Ma, Yuyang Tang, Wenfei Yang, Tianzhu Zhang, Xu Zhou, and Feng Wu. Unisot: a unified framework for multi-modality single object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

WenBei Mao, Jin Zheng, and Bo Li. Robust patch-based tracking using valid patch selection and feature fusion update. In *2014 IEEE International Conference on Image Processing (ICIP)*, pp. 4937–4941, 2014. doi: 10.1109/ICIP.2014.7026000.

Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. volume 9905, pp. 445–461, 10 2016. ISBN 978-3-319-46447-3. doi: 10.1007/978-3-319-46448-0_27.

Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 300–317, 2018.

Liang Peng, Junyuan Gao, Xinran Liu, Weihong Li, Shaohua Dong, Zhipeng Zhang, Heng Fan, and Libo Zhang. Vasttrack: Vast category visual object tracking. *Advances in Neural Information Processing Systems*, 37:130797–130818, 2024.

Zikai Song, Run Luo, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. Compact transformer tracker with correlative masked modeling. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2321–2329, 2023.

Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Chuanxin Tang, Xiyang Dai, Yucheng Zhao, Yujia Xie, Lu Yuan, and Yu-Gang Jiang. Look before you match: Instance understanding matters in video object segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2268–2278. IEEE, 2023a.

Ning Wang, Wengang Zhou, Jie Wang, and Houqiang Li. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1571–1580, 2021a.

Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 1328–1338, 2019.

Xiao Wang, Xiujun Shu, Zhipeng Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng Wu. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13763–13773, 2021b.

Xiao Wang, Jianing Li, Lin Zhu, Zhipeng Zhang, Zhe Chen, Xin Li, Yaowei Wang, Yonghong Tian, and Feng Wu. Visevent: Reliable object tracking via collaboration of frame and event flows. *IEEE transactions on cybernetics*, 54:1997–2010, 2023b.

Xingmei Wang, Guohao Nie, Jiaxiang Meng, and Zining Yan. Mimtrack: In-context tracking via masked image modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7979–7987, 2025a.

Zhenhai Wang, Lutao Yuan, Ying Ren, Sen Zhang, and Hongyu Tian. Adstrack: adaptive dynamic sampling for visual tracking. *Complex & Intelligent Systems*, 11:79, 2025b.

Xing Wei, Yifan Bai, Yongchao Zheng, Dahu Shi, and Yihong Gong. Autoregressive visual tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9697–9706, 2023.

Hao Wu, Xudong Wang, Jialiang Zhang, Junlong Tong, Xinghao Chen, Junyan Lin, Yunpu Ma, and Xiaoyu Shen. Utptrack: Towards simple and unified token pruning for visual tracking, 2026. URL https://arxiv.org/abs/2602.23734.

Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37:1834–1848, 2015. doi: 10.1109/TPAMI.2014.2388226.

Fei Xie, Chunyu Wang, Guangting Wang, Yue Cao, Wankou Yang, and Wenjun Zeng. Correlation-aware deep tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8751–8760, 2022.

Jinxia Xie, Bineng Zhong, Zhiyi Mo, Shengping Zhang, Liangtao Shi, Shuxiang Song, and Rongrong Ji. Autoregressive queries for adaptive tracking with spatio-temporal transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19300–19309, 2024.

Jinxia Xie, Bineng Zhong, Qihua Liang, Ning Li, Zhiyi Mo, and Shuxiang Song. Robust tracking via mamba-based context-aware token learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8727–8735, 2025.

Chenlong Xu, Bineng Zhong, Qihua Liang, Yaozong Zheng, Guorong Li, and Shuxiang Song. Less is more: Token context-aware learning for object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8824–8832, 2025.

Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10448–10457, 2021.

Dawei Yang, Jianfeng He, Yinchao Ma, Qianjin Yu, and Tianzhu Zhang. Foreground-background distribution modeling transformer for visual object tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10117–10127, 2023.

Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *European conference on computer vision*, pp. 341–357. Springer, 2022.

Xiaohan Yu, Jun Wang, Yang Zhao, and Yongsheng Gao. Mix-vit: Mixing attentive vision transformer for ultra-fine-grained visual categorization. *Pattern Recognition*, 135:109131, 2023.

Jiaming Zhang, Cheng Liang, Yichun Yang, Chenkai Zeng, Yutao Cui, Xinwen Zhang, Xin Zhou, Kai Ma, Gangshan Wu, and Limin Wang. Sam 2++: Tracking anything at any granularity. *arXiv preprint arXiv:2510.18822*, 2025.

Yaozong Zheng, Bineng Zhong, Qihua Liang, Zhiyi Mo, Shengping Zhang, and Xianxian Li. Odtrack: Online dense temporal token learning for visual tracking. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 7588–7596, 2024.

APPENDIX

A TRAINING DETAILS

Since ATR is introduced as a plug-in module, we follow the original training settings and evaluation protocols of each baseline tracker. Detailed configurations can be found in the corresponding papers (Lin et al., 2024; Huang et al., 2026; Wu et al., 2026; Chen et al., 2025).

We do not modify their backbone architectures, prediction heads, training objectives, or optimization strategies. ATR is only applied during template token encoding. For multi-frame trackers, ATR refines only the initial template, while subsequent dynamic templates remain unchanged.

B DATASET DETAILS

LaSOT (Fan et al., 2019) is a large-scale benchmark for long-term single-object tracking. It contains 1,400 video sequences covering 70 object categories, with 20 sequences per category. The benchmark includes various challenging factors, such as occlusion, appearance variation, deformation, and fast motion.

LaSOT$_{\text{ext}}$ (Fan et al., 2021) is an extension of LaSOT introduced in its journal version. It contains 150 additional video sequences from 15 novel object categories that do not overlap with the original 70 categories, enabling evaluation on unseen target categories.

GOT-10k (Huang et al., 2019) is a large-scale single-object tracking benchmark designed to evaluate tracker generalization. It covers a wide range of object categories and motion patterns and adopts a one-shot evaluation protocol, where object classes in the test set are separated from those used for training.

TrackingNet (Muller et al., 2018) is a large-scale benchmark for single-object tracking collected from videos in the wild. It contains a diverse range of target categories, motion patterns, and real-world scenarios, providing a comprehensive benchmark for evaluating the generalization and robustness of visual trackers.

COCO (Lin et al., 2014) is a large-scale image dataset for object detection, segmentation, and recognition. It contains diverse object categories in complex natural scenes with instance-level annotations, providing rich object appearance and contextual variations for model training.

UAV123 (Mueller et al., 2016) is a widely used aerial tracking benchmark consisting of 123 video sequences captured from low-altitude UAVs.

OTB100 (Wu et al., 2015) is another widely used tracking benchmark that includes 100 video sequences across a variety of object categories and tracking scenarios.

VastTrack (Peng et al., 2024) is a large-scale tracking benchmark designed to evaluate trackers in diverse and complex real-world scenarios. It covers a broad range of object categories and challenging conditions, including background clutter, occlusion, scale variation, and substantial appearance changes.

TNL2K (Wang et al., 2021b) (Tracking in the Wild 2K) is a large-scale benchmark designed for tracking in unconstrained real-world environments. It contains 2,000 video sequences with diverse target categories and challenging factors, such as occlusion, fast motion, deformation, illumination variation, and background clutter.

VOT-RGBD22 (Kristan et al., 2022) is a multi-modal tracking benchmark that provides synchronized RGB and depth information. It is designed to evaluate RGB-D trackers under challenging conditions where depth information can complement conventional visual appearance cues.

LasHeR (Li et al., 2021) is a large-scale RGB-T tracking benchmark containing paired visible and thermal infrared sequences. It covers diverse target categories and challenging conditions, enabling comprehensive evaluation of trackers that exploit complementary information from visible and thermal modalities.

VisEvent (Wang et al., 2023b) is a multi-modal tracking benchmark combining conventional RGB frames with event-camera observations. Event signals provide high temporal resolution and are particularly useful under challenging conditions such as fast motion and illumination changes, making VisEvent suitable for evaluating RGB-event tracking methods.