Rethinking Temporal Modeling in Visual Object Tracking via Decoupled Auxiliary Supervision

Dailing Zhang[1,2], Shiyu Hu[3], Honghao Fu[4], Xiaokun Feng[1,2], Yipei Wang[5], Kang Hao Cheong[3,†], and Kaiqi Huang[1,2,†]

[1] School of Artificial Intelligence, UCAS, China
[2] The Key Laboratory of Cognition and Decision Intelligence for Complex Systems, CASIA, China
[3] School of Physical and Mathematical Sciences, NTU, Singapore
[4] University of Queensland, Australia
[5] School of Automation, Southeast University, China
zhangdailing2023@ia.ac.cn,shiyu.hu@ntu.edu.sg,honghao.fu@uq.edu.au, fengxiaokun2022@ia.ac.cn,wyppyw@seu.edu.cn, kanghao.cheong@ntu.edu.sg,kqhuang@nlpr.ia.ac.cn

Abstract. Visual object tracking relies on modeling cross-frame dynamics, and recent approaches employ learnable temporal tokens to encode historical information. However, our analysis reveals that these tokens are often marginalized by dominant spatial features during joint optimization. This imbalance leads to shortcut learning where the network over-relies on current-frame appearance, causing temporal representation collapse. To address this, we propose **DASTrack**, a framework featuring **D**ecoupled **A**uxiliary **S**upervision (DAS). By masking the templates during an auxiliary forward pass, DAS compels the network to localize targets using only temporal tokens and thus encode robust cross-frame priors. This mechanism is applied strictly during training and discarded at inference to preserve the original architecture and computational efficiency. DASTrack yields consistent and interpretable average performance gains across multiple backbones. Our findings demonstrate that the primary bottleneck in temporal tracking stems from biased learning dynamics rather than architectural design, offering a novel training paradigm for eliciting latent temporal capabilities. The code and models will be released at: https://github.com/ZhangDailing8/DASTrack.

Keywords: visual object tracking · temporal representation learning · multi-task learning

1 Introduction

Visual Object Tracking (VOT) [28, 48] aims to continuously localize a target throughout a long video sequence and serves as a fundamental task in video

† Corresponding authors

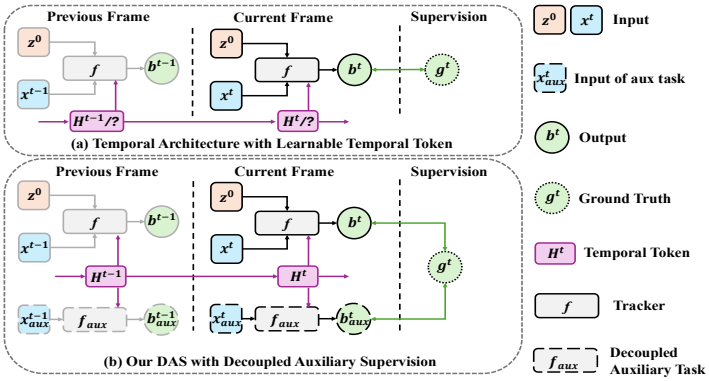

Fig. 1: Temporal learning mechanism in visual object tracking. (a) In standard joint training, dense spatial supervision dominates optimization, allowing localization from appearance and causing the temporal token to receive weak and ineffective gradients, resulting in a spatial shortcut and ineffective cross-frame representation. (b) DASTrack introduces Decoupled Auxiliary Supervision to provide a dedicated learning signal for the temporal token, establishing an independent temporal learning pathway that breaks representation competition and restores temporal modeling.

understanding. With the advent of Vision Transformers (ViTs) [11], tracking frameworks have gradually evolved from siamese two-branch matching architectures [3,29,60] to one-stream attention paradigms [10,51], where feature extraction and relational reasoning are jointly modeled within a unified representation space. Although this design substantially improves target–background discriminability, long-term tracking remains challenging in the presence of dynamic challenges like occlusion and deformation [13, 23, 46]. Reliable tracking in such scenarios requires consistent utilization of historical information across frames. Consequently, recent approaches introduce temporal tokens or historical feature propagation mechanisms to explicitly model temporal dynamics [1,26,42,49,59].

However, we find that the actual contribution of these temporal structures [1, 42, 49, 59] to model decisions is much weaker than expected. We first analyze model behavior directly. As illustrated in Fig. 1(a), freezing or perturbing the temporal tokens during inference leads to almost no performance degradation and trackers can still localize the target stably. To further examine this phenomenon, we conduct a controlled preliminary study (Tab. 1) by intervening in cross-frame information. When more accurate temporal cues are provided, performance improves substantially (nearly 10%), whereas disrupting the existing temporal information causes negligible change in the original tracker. This asymmetric result indicates that current trackers do not lack temporal modeling capacity; rather, they fail to actually utilize temporal information in training.

This observation reveals an optimization issue. Under joint training, dense spatial supervision provides a direct and stable localization signal, allowing the network to converge to a shortcut solution that relies primarily on single-frame

appearance matching. In contrast, temporal tokens participate in optimization indirectly, and their gradients are systematically suppressed. This leads to an implicit *spatial domination*: although the architecture contains temporal modeling components, the decision-making process remains dominated by spatial features. This explains why perturbing temporal tokens barely affects predictions.

Motivated by this finding, we propose **DASTrack** (Decoupled Auxiliary Supervision Tracking). As shown in Fig. 1(b), we introduce a decoupled auxiliary supervision during training that prevents the model from directly relying on template appearance, forcing localization to depend on temporal tokens and thereby providing an independent learning signal for the temporal pathway. The mechanism is applied only during training and introduces no additional modules or computational overhead during inference [51, 57, 59]. Experiments show that DASTrack consistently improves performance across different backbones, yielding average performance gains of approximately 0.5% to 0.8%. More importantly, when temporal tokens are perturbed at inference, the tracker exhibits a consistent and controlled performance degradation aligned with the spatial-only baseline, demonstrating that temporal information genuinely participates in the decision process rather than remaining a redundant representation.

In summary, the contributions of this work are threefold: (1) We reveal, through behavioral intervention and quantitative evidence presented in Sec. 3, that temporal tokens in existing trackers are marginalized by spatial features under joint optimization. (2) We propose the DAS training paradigm in Sec. 4, which restores temporal representation learning without modifying the network architecture or inference cost. (3) We validate its effectiveness and generalization across multiple benchmarks and backbones in Sec. 5, and further show from a model-behavior perspective that the proposed mechanism changes the decision dependency of the tracker.

2 Related Work

2.1 Temporal Modeling in Visual Object Tracking

Early VOT methods commonly formulated the task as an appearance matching problem based on the initial template [3, 4, 7, 8, 51]. In short-term tracking scenarios, target motion is typically smooth and appearance variations are limited, allowing trackers to localize the target using features from the current frame alone [24, 38]. However, in longer video sequences, dynamic challenges like occlusions and deformations frequently occur, and template matching strategies often fail [13, 23]. This suggests that reliable long-term tracking fundamentally requires the consistent utilization of historical information across frames [58].

To address this, subsequent studies introduced explicit temporal modeling mechanisms. One line of work performs online template updating to adapt to appearance changes, but the update strategies often rely on heuristic rules and involve a trade-off between stability and computational overhead [5, 9, 50]. Another line leverages historical trajectories or geometric motion priors to enforce

temporal continuity, yet geometric cues alone are insufficient to represent complex semantic variations [32, 47]. Overall, these approaches incorporate cross-frame information mainly at the input or data level, while the model's decision process still largely relies on current-frame appearance matching.

With the adoption of ViT [11] in tracking, recent approaches introduce learnable temporal tokens to propagate contextual information across frames within a unified feature space [1, 2, 6, 14, 59]. This design provides an efficient architectural mechanism for temporal modeling and has become a core component of modern tracking frameworks. Existing studies [15, 16, 26, 30, 49] primarily focus on designing stronger temporal modules or improving information propagation strategies, and their effectiveness is typically evaluated indirectly through performance improvements.

2.2 Shortcut Learning

Shortcut learning describes the tendency of deep models to rely on superficial yet easily exploitable cues, rather than task-relevant signals that generalize under distribution shift [17, 43, 54]. Prior work has shown that such shortcut solutions are pervasive in modern recognition systems and are often preferred not only because they are predictive, but also because they are easier for the model to extract during optimization [21, 52]. In vision tasks, this behavior typically manifests as an over-reliance on appearance, texture, or contextual background, instead of more robust semantic or structural cues [18, 41, 45]. As a result, models may achieve strong in-distribution performance while exhibiting limited robustness when challenges occur.

3 Temporal Information in Visual Object Tracking

Long-term VOT has recently shifted from static current-frame matching [4, 51] to temporal paradigms [9, 32, 59] that explicitly model cross-frame relations. To mitigate accumulated errors from occlusion or deformation, modern trackers often incorporate learnable temporal tokens or memory states to aggregate historical context [1, 42, 49, 59]. Ideally, these modules provide critical supplementary cues to maintain target consistency when appearance matching fails. However, the architectural presence of a temporal channel does not guarantee its actual utilization during the decision process. Existing literature broadly assumes that integrating these modules inherently confers temporal reasoning capabilities, yet this remains empirically unverified. Therefore, we investigate a fundamental question: *do existing trackers truly exploit temporal information for localization?* Through controlled intervention experiments, we systematically narrow the interpretation space and summarize our findings in Tab. 1.

3.1 Temporal Trackers

Temporal Trackers. In a standard VOT formulation, given an initial target template, the model continuously predicts the target location within the

Table 1: Performance comparison on the LaSOT benchmark under original and modified settings. To evaluate the impact of temporal cues, we introduce two modifications on non-learnable temporal trackers: (#1) cropping the online template using GT coordinates, and (#2) centering the search region strictly on GT coordinates. The * denotes the static baseline variants that completely omit temporal modules. Δ represents the average metric of the modified setting minus that of the original setting.

Tracker	LaSOT (Original)			LaSOT (Modified)			Δ
	AUC	P_{Norm}	P	AUC	P_{Norm}	P	
More accurate temporal information							
SUTrack [4] (# 1)	73.1	83.3	80.4	82.1	92.2	91.6	9.7
SUTrack [4] (# 2)	-	-	-	79.8	91.5	88.4	7.6
OSTrack [51] (# 1)	69.9	79.8	76.2	73.9	84.6	81.9	4.8
OSTrack [51] (# 2)	-	-	-	77.9	90.0	86.7	9.6
Perturb temporal information in official checkpoint							
ARTrackV2 [1]	71.4	80.2	77.4	71.2	80.6	77.3	0.03
EVPTrack [42]	70.2	80.7	77.0	70.1	80.5	76.8	-0.17
AQATrack [49]	71.2	81.7	78.4	71.4	81.8	78.6	0.17
ODTrack [59]	72.9	82.8	80.1	72.9	82.9	80.3	0.10
Perturb temporal information in reproduce checkpoint							
EVPTrack [42]	69.9	80.2	76.5	70.0	80.3	76.4	0.03
EVPTrack*	69.9	80.1	76.7	-	-	-	0.03
AQATrack [49]	70.6	80.7	77.5	70.7	80.8	77.6	0.10
AQATrack*	71.4	81.7	78.6	-	-	-	0.97
ODTrack [59]	72.1	82.0	78.8	71.8	81.7	78.5	-0.30
ODTrack*	71.8	81.8	78.7	-	-	-	-0.20

search regions of subsequent frames. Current state-of-the-art (SOTA) methods [4, 19, 26, 31] predominantly adopt a one-stream architecture. Specifically, multiple target templates $\boldsymbol{T}_i \in \mathbb{R}^{H_z \times W_z \times 3}$ and the search region $\boldsymbol{S} \in \mathbb{R}^{H_x \times W_x \times 3}$ are tokenized into sequences $\boldsymbol{Z}_i \in \mathbb{R}^{N_T \times d}$ and $\boldsymbol{X} \in \mathbb{R}^{N_S \times d}$ respectively, and then concatenated. Here $N_T = \frac{H_z W_z}{P^2}$ and $N_S = \frac{H_x W_x}{P^2}$ represent the corresponding token count. A ViT encoder $\boldsymbol{\Phi}$ processes this sequence to extract joint representations $\boldsymbol{F}$ [11, 20, 55]. The output tokens corresponding to the search region are subsequently fed into a prediction head $\boldsymbol{\Psi}$ for prediction.

Unlike traditional static trackers, temporal trackers introduce a compact set of learnable temporal tokens $\boldsymbol{H}$ into this sequence to capture and propagate historical context across frames. These tokens are explicitly designed to provide robust temporal priors during appearance variations or severe occlusions, ensuring the model does not rely exclusively on intra-frame spatial matching. This

temporal propagation process is mathematically defined as follows:

$$\boldsymbol{I} = \text{Concat}(\boldsymbol{H}, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N, \boldsymbol{X}), \quad \boldsymbol{F} = \boldsymbol{\Phi}(\boldsymbol{I}), \tag{1}$$

$$[\boldsymbol{H}', \boldsymbol{Z}'_1, \ldots, \boldsymbol{Z}'_N, \boldsymbol{X}'] = \text{Split}(\boldsymbol{F}), \tag{2}$$

$$\boldsymbol{B} = \boldsymbol{\Psi}(\boldsymbol{X}'), \tag{3}$$

where $\boldsymbol{B}$ denotes the predicting bounding box.

If the network genuinely utilizes historical context, perturbing the temporal tokens should noticeably degrade tracking performance. Conversely, unaffected performance would indicate that current-frame spatial matching dominates the localization process. Driven by this testable hypothesis, we begin our empirical analysis to investigate the true utility of temporal information.

3.2 The Temporal Utilization Gap

Observation 1: Accurate historical information significantly enhances tracking performance. To characterize the upper bound of an ideal temporal prior, we conduct inference-time interventions on baseline trackers lacking learnable temporal modules. Since their implicit historical context relies on previous predictions and is vulnerable to error accumulation, we implement two oracle settings: cropping the online templates using the ground truth (GT) bounding box and centering the search region precisely on the GT coordinates. As demonstrated in Tab. 1, both configurations deliver substantial performance gains, such as the AUC increase from 73.1% to 82.1% for SUTrack [4] and 69.9% to 77.9% for OSTrack [51]. These findings suggest that cross-frame information holds intrinsic value, as long-term tracking performance improves significantly when historical context is effectively utilized. Consequently, temporal information should be viewed as a critical cue for stability rather than merely an auxiliary signal. However, directly perturbing the internal temporal representations within the model reveals a completely contrasting phenomenon.

Observation 2: Perturbing temporal tokens during inference causes negligible performance drops. For trackers equipped with learnable temporal tokens, we maintain all architectural components and exclusively perturb the temporal tokens during the inference phase. The results in Tab. 1 indicate that the fluctuation in metrics is negligible, typically around 0.2%. For example, the performance of models such as ARTrackV2 [1], EVPTrack [42], AQATrack [49], and ODTrack [59] remains almost identical. As the temporal token serves as the sole explicit pathway for conveying historical information, this phenomenon cannot be simply attributed to model robustness. A more convincing explanation is that the network fails to genuinely rely on this temporal channel for localization. Instead, its predictions are predominantly decided by spatial appearance matching between the target template and the current search region. Consequently, while a cross-frame information pathway exists architecturally, the decision process effectively bypasses it. Another possible explanation is that historical context may be redundantly encoded within other feature channels. To eliminate this possibility, we introduce further structural control experiments.

Observation 3: Performance remains comparable when the temporal module is removed. Repeating the perturbation on reproduced checkpoints yields consistent conclusions. Furthermore, training a static baseline without the temporal module (denoted with an asterisk *) achieves performance closely matching the temporal-aware variant. For instance, the metrics of EVPTrack* are essentially identical to those of the original EVPTrack, and ODTrack* similarly exhibits no significant degradation. This indicates that the network does not implicitly encode historical context through alternative feature pathways. Instead, the model naturally minimizes its reliance on temporal representations during the optimization process.

Synthesizing these three observations reveals a paradox. Providing accurate historical information significantly boosts tracking performance, yet corrupting or removing the internal temporal representations leaves the predictions virtually unchanged. This implies a fundamental issue: while temporal information is undeniably valuable, the model fails to learn how to utilize it effectively.

3.3 Optimization Mechanism Analysis

The observed paradox likely stems from the joint optimization mechanism rather than structural limitations. As standard tracking objectives can often be resolved directly through high-resolution spatial appearance matching, the network rapidly converges toward this shortcut solution. The high performance of trackers without temporal modules, such as OSTrack [51] and SUTrack [4], further supports this observation. During this process, the strong gradient signals from spatial features dominate the optimization, systematically marginalizing the weaker, indirect supervision provided by temporal tokens. We term this optimization bias *spatial domination*.

Consequently, the core limitation of existing trackers is not an inadequate architecture, but a flawed training dynamic that ignores temporal modeling. Restoring this modeling capacity requires shifting the training focus to enforce a genuine reliance on historical context, which serves as the primary motivation for our Decoupled Auxiliary Supervision framework.

4 Method

4.1 Overall Architecture

Based on the analysis in Sec. 3, we aim to reform the training paradigm rather than designing a bespoke tracker architecture. This ensures the network effectively exploits temporal information while fully preserving the original inference pipeline. To this end, we propose DASTrack, centered around a novel Decoupled Auxiliary Supervision (DAS) framework.

As illustrated in Fig. 2, the standard input of temporal-aware trackers [2, 59] comprises target templates, the current search region, and temporal tokens. During training, we explicitly duplicate the tokenized search region to construct an

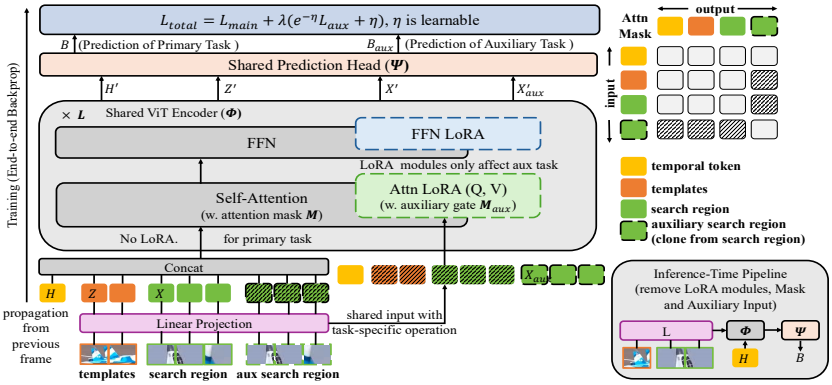

Fig. 2: Overview of DAS. Left: The unified input sequence, comprising temporal, template, search, and auxiliary search tokens, is processed by a shared ViT encoder. To achieve structural decoupling, task specific LoRA modules are integrated into self-attention and FFN layers. These modules exclusively update the auxiliary search tokens to capture temporal dynamics without interfering with the spatial feature learning of the primary branch. The entire framework is optimized end to end using an adaptive loss L_{total}. **Right**: Customized attention mask regulates the interaction between tokens. Masked regions (hatched boxes) prevent primary tokens from accessing auxiliary search tokens to eliminate information leakage. Conversely, auxiliary tokens are permitted to access temporal tokens and themselves to facilitate learning signal propagation and reinforce temporal representation modeling.

auxiliary input. This cloned region is concatenated with the original sequence and processed by the shared ViT encoder $\boldsymbol{\Phi}$. A shared prediction head $\boldsymbol{\Psi}$ subsequently generates the final tracking results. This process is formulated as:

$$\boldsymbol{I} = \text{Concat}(\boldsymbol{H}, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N, \boldsymbol{X}, \boldsymbol{X}_{aux}), \quad \boldsymbol{F} = \boldsymbol{\Phi}(\boldsymbol{I}), \tag{4}$$

$$[\boldsymbol{H}', \boldsymbol{Z}'_1, \ldots, \boldsymbol{Z}'_N, \boldsymbol{X}', \boldsymbol{X}'_{aux}] = \text{Split}(\boldsymbol{F}), \tag{5}$$

$$\boldsymbol{B} = \boldsymbol{\Psi}(\boldsymbol{X}'), \boldsymbol{B}_{aux} = \boldsymbol{\Psi}(\boldsymbol{X}'_{aux}). \tag{6}$$

Consequently, the network jointly optimizes two distinct localization objectives. The primary task performs standard template matching using temporal token, visible templates and the search region. Conversely, the auxiliary task is strictly restricted to localizing the target using only the temporal token and the auxiliary search region.

As the auxiliary branch lacks access to the templates, the model is compelled to compress comprehensive target priors into the temporal tokens. This mechanism transforms the temporal tokens from a marginalized historical cache into an indispensable representation carrier. Finally, the entire network is trained end-to-end using an adaptively weighted joint loss function.

4.2 Decoupled Auxiliary Task

Standard temporal trackers [42, 49, 59] suffer from spatial matching bias, where the network over-relies on the template and marginalizes historical context. To disrupt this sub-optimal path, we introduce a restricted auxiliary localization task. In this branch, the network must localize the target using only the current auxiliary search region and temporal tokens. By strictly isolating the auxiliary branch from the target templates, we compel the model to encode essential localization priors and appearance variations into the temporal tokens.

During training, we sample multiple templates and a sequence of search frames. This enables the temporal tokens to propagate through the video sequence and integrate information from both historical search regions and multiple templates. Both the primary and auxiliary branches share a common ViT encoder [11, 55] and prediction head [26, 51]. This competitive dynamic during joint optimization forces the model to balance explicit spatial matching with robust temporal representations, ensuring that temporal cues are genuinely utilized during the decision-making process.

4.3 Auxiliary Task Mask

However, naively duplication of the search region may introduce unintended information leakage, as the auxiliary branch could access template features through the self-attention mechanism [44] and bypass the intended temporal reasoning. To ensure rigorous isolation of the auxiliary flow, we introduce a customized attention mask within the encoder, as shown on the right side of Fig. 2. The masking matrix $\boldsymbol{M}$ and the auxiliary gate $\boldsymbol{M}_{aux}$ are formulated as follows:

$$\boldsymbol{M}_{i,j} = \begin{cases} -\infty, & \text{if } i \in (\boldsymbol{H} \cup \boldsymbol{X} \cup \boldsymbol{Z}_i) \text{ and } j \in \boldsymbol{X}_{aux}, \\ -\infty, & \text{if } i \in \boldsymbol{X}_{aux} \text{ and } j \in (\boldsymbol{X} \cup \boldsymbol{Z}_i), \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

$$(\boldsymbol{M}_{aux})_i = \begin{cases} 1, & \text{if } i \in \boldsymbol{X}_{aux}, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

This design enforces strict routing rules. Tokens from the primary branch are prohibited from attending to the auxiliary search region. Conversely, auxiliary search tokens are restricted to interacting only with the temporal tokens and themselves. By severing the spatial matching shortcut, this structural constraint forces the model to extract localization information exclusively from the temporal tokens, ensuring they encode robust and comprehensive temporal priors.

4.4 LoRA-Based Structural Decoupling

To mitigate potential gradient conflicts during joint optimization in the shared backbone [27, 53], we employ LoRA [22, 33, 34, 56] as a structural decoupling

mechanism rather than for parameter efficiency. While the primary task continues to optimize the shared backbone weights, the auxiliary task updates separate low-rank residual branches. We introduce a token-level gating mechanism $\boldsymbol{M}_{aux}$ to ensure that LoRA updates exclusively affect auxiliary search region tokens. This strategy allows the branches to operate in distinct parameter subspaces, where the primary branch maintains stable spatial matching while the auxiliary branch develops robust temporal representations.

This deployment ensures that the learning of temporal modeling remains a strictly constrained process without disrupting the spatial localization capacity. The adaptation in the attention mechanism is formulated as:

$$\boldsymbol{Q} = \boldsymbol{W}_q\boldsymbol{E} + \boldsymbol{B}_q\boldsymbol{A}_q\boldsymbol{E} \odot \boldsymbol{M}_{aux}, \quad \boldsymbol{K} = \boldsymbol{W}_k\boldsymbol{E}, \tag{9}$$

$$\boldsymbol{V} = \boldsymbol{W}_v\boldsymbol{E} + \boldsymbol{B}_v\boldsymbol{A}_v\boldsymbol{E} \odot \boldsymbol{M}_{aux}, \tag{10}$$

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \boldsymbol{V}(\text{Softmax}(\frac{\boldsymbol{Q}^\top \boldsymbol{K}}{\sqrt{d}} + \boldsymbol{M})). \tag{11}$$

Similarly, the adaptation within the Feed-Forward Network (FFN) is:

$$\boldsymbol{H}_1 = \boldsymbol{W}_1\boldsymbol{E} + \boldsymbol{B}_1\boldsymbol{A}_1\boldsymbol{E} \odot \boldsymbol{M}_{aux}, \quad \boldsymbol{H}_2 = \boldsymbol{W}_2\boldsymbol{E} + \boldsymbol{B}_2\boldsymbol{A}_2\boldsymbol{E} \odot \boldsymbol{M}_{aux} \tag{12}$$

$$\boldsymbol{Y} = \boldsymbol{W}_3\boldsymbol{H}_{gate} + \boldsymbol{B}_3\boldsymbol{A}_3\boldsymbol{H}_{gate} \odot \boldsymbol{M}_{aux}, \quad \text{where } \boldsymbol{H}_{gate} = \sigma(\boldsymbol{H}_1) \odot \boldsymbol{H}_2 \tag{13}$$

and σ is the activation function. Specifically, we utilize HiViT [55] in our primary implementation, though the framework generalizes to standard ViT backbones [11], as detailed in Sec. 5.3.

4.5 Optimization and Inference

Following prior work [49, 59], we perform video sequence-level training based on the DAS framework, optimizing the network in an end-to-end manner to instantiate DASTrack.

Loss. Regarding the tracking predictions, following OSTrack [51], we employ a weighted focal loss [35] for classification, an ℓ_1 loss and a generalized IoU loss [40] for bounding box regression. Additionally, a dedicated quality loss is applied to supervise the tracking quality estimation, following ARTrackV2 [1]. Both the primary and auxiliary tasks are optimized using identical loss function:

$$L = L_{cls} + \lambda_1 L_1 + \lambda_2 L_{GIoU} + \lambda_3 L_{Quality}. \tag{14}$$

To further mitigate potential negative impact of the auxiliary task on the primary tracking objective, we introduce an adaptive weighting mechanism grounded in Maximum Likelihood Estimation (MLE) [27]. Specifically, we formulate the task uncertainty (η) of the auxiliary task as a dynamically learnable parameter. By optimizing the negative log-likelihood, the network autonomously discovers the optimal balance between the simultaneous tasks, thereby obviating the need for exhaustive manual grid searches for loss hyperparameters. This adaptively weighted loss function and total loss function are formulated as:

$$\tilde{L}_{aux} = e^{-\eta} L_{aux} + \eta, \quad L_{total} = L_{main} + \lambda \tilde{L}_{aux}. \tag{15}$$

Table 2: Benchmarking our tracker on long-term tracking datasets. Note that all compared methods feature medium-scale architectures for fair comparison. **Bold** indicates the best results and underline indicates the second-best.

Tracker	LaSOT [13]			$\text{LaSOT}_{\text{ext}}$ [12]			TNL2k [46]		VideoCube [23]	
	AUC	P_{Norm}	P	AUC	P_{Norm}	P	AUC	P	SR	P
DASTrack	**73.4**	**84.8**	**81.5**	**53.6**	**64.9**	**61.1**	64.6	**69.9**	**71.3**	**66.4**
SiamFC_{255} [3]	33.6	42.0	33.9	23.0	31.1	26.9	29.5	28.6	6.1	3.1
GlobalTrack [25]	52.1	-	52.7	-	-	-	40.5	38.6	45.5	29.7
TransT_{256} [8]	64.9	73.8	69.0	44.8	52.3	52.5	50.7	51.7	53.9	44.7
OSTrack_{256} [51]	69.1	78.7	75.2	47.4	57.3	53.3	54.3	-	58.3	46.7
ARTrack_{256} [47]	70.4	79.5	76.6	46.4	56.5	52.3	57.5	-	59.9	51.6
LoRAT_{224} [34]	71.7	80.9	77.3	50.3	61.6	57.1	58.8	61.3	-	-
ODTrack_{384} [59]	73.2	83.2	80.6	52.4	63.9	60.1	60.9	-	69.1	62.2
AQATrack_{256} [49]	71.4	81.9	78.6	51.2	62.2	58.9	57.8	59.4	65.6	57.6
ARTrackv2_{256} [1]	71.6	80.2	77.2	50.8	61.9	57.7	59.2	-	56.8	49.8
SUTrack_{224} [4]	73.2	83.4	80.5	53.1	64.2	60.5	**65.0**	67.9	67.9	62.1
LoRATv2_{224} [33]	72.0	81.3	77.9	-	-	-	59.6	62.7	-	-
Some Trackers with Higher Resolution										
OSTrack_{384} [51]	71.1	81.1	77.6	50.5	61.3	57.6	55.9	56.7	58.3	47.6
ARTrack_{384} [47]	72.6	81.7	79.1	51.9	62.0	58.5	-	-	-	-
LoRAT_{378} [34]	72.9	82.0	79.1	51.7	**65.1**	59.2	58.6	60.8	-	-
ODTrack_{384} [59]	73.2	83.2	80.6	52.4	63.9	60.1	60.9	-	69.1	62.2
AQATrack_{384} [49]	72.7	82.9	80.2	52.7	64.2	60.8	59.3	62.3	69.1	63.2
SUTrack_{384} [4]	74.4	83.9	81.9	52.9	63.6	60.1	65.6	69.3	69.7	66.9
LoRATv2_{378} [33]	74.3	83.6	80.9	-	-	-	60.9	64.9	-	-
DASTrack_{384}	**74.6**	**84.6**	**82.9**	**53.5**	64.2	**61.0**	**66.4**	**73.0**	**73.6**	**71.9**

Inference. During the inference phase, both the auxiliary search region inputs and the task-specific LoRA modules are removed. This elegant decoupling ensures that the original baseline architecture remains structurally unaltered, introducing minimal additional computational overhead.

5 Experiments

5.1 Implementation Details

Model Variants. To validate generalizability, we instantiate DASTrack with a SUTrack-pretrained [4] HiViT-Base. Both variants uniformly adopt crop factors of 2 and 4 for the templates and search region respectively, alongside 3 reference frames ($N = 3$) for spatio-temporal modeling. Resolution-wise, DASTrack processes 112×112 templates and 224×224 search regions. For DAS training, LoRA modules (default rank 64) are strategically injected into the FFN, the query and value projection matrices of the multi-head self-attention mechanism.

Table 3: Benchmarking our tracker on short-term tracking datasets. Where * denotes only trained on GOT-10k.

Tracker	TrackingNet [38] AUC	GOT-10k* [24] AO
OSTrack_{256} [51]	83.1	71.0
ARTrack_{256} [47]	84.2	73.5
LoRAT_{224} [34]	83.5	72.1
ODTrack_{384} [59]	<u>85.1</u>	<u>77.0</u>
LoRATv2_{224} [33]	-	74.7
DASTrack	**85.3**	76.5
DASTrack_{384}	**86.6**	**79.0**

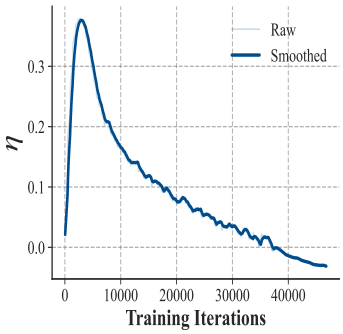

Fig. 3: Evolution of auxiliary task uncertainty (inversely proportional to task weight).

Table 4: Ablation of adaptive weight.

Weight	LaSOT		
	AUC	P_{Norm}	P
adaptive	73.4	84.8	81.5
fixed	73.2	84.2	81.2

Table 5: Ablation of visible information in auxiliary task.

Visual Information	LaSOT			LaSOT(Modified)			Δ
	AUC	P_{Norm}	P	AUC	P_{Norm}	P	
temporal token	73.4	84.8	81.5	72.6	83.8	80.6	-0.90
+ one template	73.0	84.2	81.2	72.9	84.0	81.0	-0.17
+ all templates	72.6	83.6	80.4	72.8	84.0	80.8	0.33

Training. Following standard protocols, our model is trained on LaSOT [13], GOT-10k [24], TrackingNet [38], COCO [36], and VastTrack [39]. For the evaluation of the GOT-10k test set, we strictly follow its one-shot protocol by training our models on the complete GOT-10k training split. Training inputs consist of a 7-frame sequence (3 templates, 4 search regions). By concatenating these frames, temporal tokens effectively propagate inter-frame context and gradient flows. We optimize the network using AdamW [37] (weight decay 1×10^{-4}) for 100 epochs of 60,000 sequences each, decaying the learning rates by a factor of 10 at epoch 80. Initial learning rates are differentially set: 4×10^{-5} for the backbone, 4×10^{-4} for task-specific components, and 4×10^{-3} for learnable task weights. Training is conducted on 4 NVIDIA A40 GPUs with a batch size of 32.

Inference. During inference, videos are processed in 4-frame clip to align with training. We maintain an ensemble of one static and two dynamic templates. The dynamic templates are periodically updated using high confidence predictions saved in a memory bank. Finally, a standard hamming window penalizes large displacements on the classification response map.

5.2 State-of-the-Art Comparisons

We evaluate DASTrack on long-term and short-term tracking benchmarks and provide corresponding visualization results in Fig. 4. Additionally, we provide details on the computational efficiency of DASTrack in Tab. 9.

Table 6: Ablation study of LoRA parameters.

Optimization Settings	LaSOT AUC	LaSOT P_{Norm}	LaSOT P
Sequential	72.7	84.0	80.8
Joint			
w/o LoRA	73.2	84.3	81.0
rank=16	72.6	83.7	80.6
rank=32	73.0	83.8	81.0
rank=64	73.4	84.8	81.5
rank=128	72.7	83.8	80.8

Table 7: Ablation study of temporal modeling. Compares our method (DAS), standard joint training with temporal token (Temporal), and the spatial-only framework (Spatial).

Training Method	LaSOT AUC	LaSOT P_{Norm}	LaSOT P	LaSOT(Modified) AUC	LaSOT(Modified) P_{Norm}	LaSOT(Modified) P	Δ
HiViT							
DAS	73.4	84.8	81.5	72.6	83.8	80.6	-0.90
Temporal	73.0	84.1	80.9	73.0	84.1	80.9	-0.03
Spatial	72.8	83.9	80.7	-	-	-	-
ViT							
DAS	71.4	81.8	78.7	71.0	81.3	78.1	-0.50
Temporal	71.0	81.2	77.9	70.9	81.0	77.8	-0.13
Spatial	71.0	81.4	78.0	-	-	-	-

Table 8: Gradient diagnostics and inference-time temporal utilization. R_H measures the temporal-to-spatial gradient ratio, G_H^{aux} ratio denotes the auxiliary contribution to temporal-token gradients, and Norm-TAR denotes temporal attention normalized by the temporal-token proportion. Drop is the performance change under temporal perturbation.

Method	$R_H \uparrow$	$G_H^{\text{total}} \uparrow$	G_H^{aux} ratio	Norm-TAR	Drop
Temporal	0.165	0.00115	–	2.3	-0.03
DAS	0.624	0.00504	27.6%	14.4	-0.90

Table 9: Details of DASTrack model variants.

Trackers	Params (M)	FLOPs (G)	Speed *fps*
SUTrack_{256}	71	26	34.0
DASTrack_{256}	71	26	34.0
SUTrack_{384}	77	71	28.1
DASTrack_{384}	77	71	28.1

Long-Term Tracking. As shown in Tab. 2, DASTrack establishes a SOTA across LaSOT, $\text{LaSOT}_{\text{ext}}$, and VideoCube, notably achieving an unprecedented 71.3% SR on the latter. It also delivers highly competitive results on TNL2k with the best precision (69.9%) and second best AUC (64.6%). These results demonstrate that DASTrack maintains consistent localization stability in long-term sequences characterized by severe dynamic challenges.

Short-Term Tracking. On short-term tracking benchmarks shown in Tab. 3, DASTrack also remains highly competitive. It achieves the highest AUC of 85.3% on TrackingNet and the second highest AO of 76.5% on GOT-10k. It is noteworthy that the top-ranking ODTrack_{384} utilizes a higher input resolution, whereas our method delivers comparable robustness under a unified resolution setting. This indicates that DASTrack is not a specialized technique limited to long-term scenarios but a robust temporal modeling paradigm that generalizes naturally to short-term tracking without increasing inference complexity.

5.3 Ablation Experiments

We conduct experiments to investigate the key characteristics of the DAS paradigm, including temporal modeling, shortcut learning behavior, structural decoupling, and optimization dynamics. Unless otherwise specified, all analyses are evaluated on LaSOT. Default configurations are highlighted in gray.

Effectiveness & Generalizability. To analyze the role of temporal modeling, we compare DAS against a spatial-only baseline (*Spatial*) and a standard joint training variant (*Temporal*). Tab. 7 shows that the *Temporal* variant yields negligible gains over the *Spatial* baseline for both HiViT [55] and ViT [11] backbones, indicating a failure to utilize temporal cues despite their architectural presence. In contrast, DAS achieves consistent improvements of 0.5% to 0.8%. These results confirm that the performance boost stems from the modified optimization path rather than increased capacity, as decoupled supervision compels the network to encode and leverage vital inter-frame dependencies.

Robustness & Interpretability. To evaluate temporal dependency, we perturb temporal tokens during inference. As shown in Tab. 7, *Temporal* variant results in negligible fluctuation, approximately -0.03%, indicating bypassed temporal cues. Conversely, DASTrack exhibits a logical performance drop, specifically -0.9% on HiViT and -0.5% on ViT. This degraded performance closely matches the *Spatial* baseline, such as 72.6% versus 72.8% AUC on HiViT. Such graceful degradation confirms that DAS compels the network to encode vital temporal priors. When these priors are corrupted, the model naturally reverts to its baseline spatial matching capacity, proving the genuine engagement of the temporal module in the decision-making process.

Auxiliary Task Visibility & Shortcut Learning. We investigate the impact of visual context by progressively exposing template tokens to the auxiliary branch. Counter-intuitively, richer visual context degrades performance, where the AUC drops from 73.4% to 72.6% in Tab. 5. This phenomenon validates our spatial domination hypothesis, suggesting that when templates are accessible, the network bypasses temporal modeling in favor of explicit spatial matching, leading to shortcut learning. Consequently, strictly masking all templates is crucial as it compels the network to compress comprehensive localization priors into the temporal tokens, enabling the formation of genuine temporal representations.

LoRA Rank & Decoupling Capability. To further understand the role of structural decoupling, we analyze the effect of the LoRA rank r in Tab. 6. Performance is constrained when $r \leq 32$ due to insufficient capacity for auxiliary information. Conversely, $r = 128$ causes a performance decline as excessive redundancy introduces task interference. Setting $r = 64$ achieves the optimal balance between representation and decoupling. Furthermore, joint optimization improves AUC by 0.5% over the sequential baseline. This indicates that the auxiliary task serves as a complementary constraint within the shared feature space rather than an isolated training phase.

Adaptive Loss Weighting & Optimization Dynamics. To mitigate auxiliary task interference, we employ an MLE-based adaptive weighting mechanism [27], formulating task uncertainty (η) as a learnable parameter. This strat-

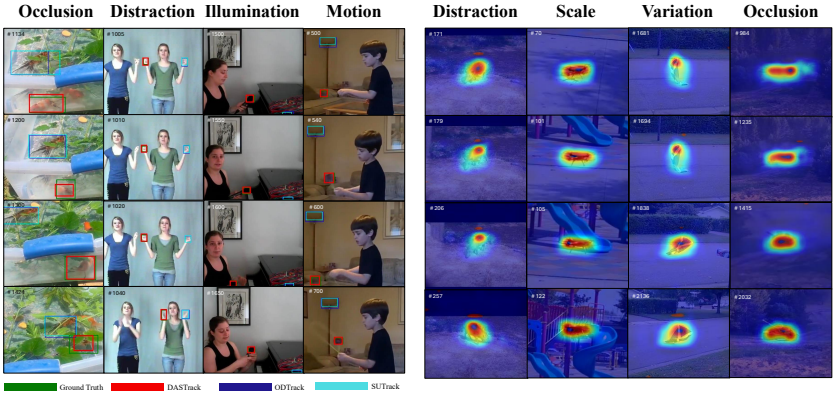

Fig. 4: Qualitative tracking results and attention visualization. Left. Visual comparison of DASTrack against existing trackers across challenging scenarios. **Right.** Attention heatmaps of the temporal token relative to the search region in DASTrack.

egy yields a 0.4% average performance gain (Tab. 4). Notably, Fig. 3 reveals a robust "suppress-then-fuse" optimization dynamic: η initially spikes to down-weight the auxiliary loss, effectively suppressing early gradient noise. As training stabilizes and uncertainty decreases, the auxiliary weight naturally recovers. This late-stage fusion allows the auxiliary supervision to enhance temporal feature generalization without compromising the primary tracking convergence.

Optimization Dynamics & Temporal Utilization. Standard joint training may favor the easier template–search shortcut, $f_\theta(Z, X_t, H_{t-1}) \approx f_\theta(Z, X_t)$, leaving temporal tokens weakly supervised. DAS removes template access in the auxiliary branch, forcing reliance on H_{t-1} and providing a direct temporal gradient. As shown in Tab. 8, DAS increases R_H from 0.165 to 0.624, with 27.6% auxiliary contribution, and raises Norm-TAR from 2.3 to 14.4. The larger perturbation drop, from -0.03 to -0.90, further confirms stronger temporal utilization during inference.

6 Conclusion

We identify a *spatial dominance* in temporal trackers where models prioritize spatial matching over temporal dependencies during joint optimization. We propose DAS, a training-time decoupled auxiliary framework that compels the network to encode robust localization priors into temporal tokens. DASTrack performs competitively on several benchmarks while introducing zero inference overhead. Perturbation experiments demonstrate that the model naturally reverts to its spatial baseline capacity when temporal cues is corrupted, which validates the genuine engagement and interpretability of the learned temporal features.

Acknowledgments. This work is supported in part by the National Science and Technology Major Project (Grant No.2022ZD0116403).

References

1. Bai, Y., Zhao, Z., Gong, Y., Wei, X.: Artrackv2: Prompting autoregressive tracker where to look and how to describe. In: CVPR. pp. 19048–19057 (2024)
2. Cai, W., Liu, Q., Wang, Y.: Spmtrack: spatio-temporal parameter-efficient fine-tuning with mixture of experts for scalable visual tracking. In: Proceedings of the computer vision and pattern recognition conference. pp. 16871–16881 (2025)
3. Cen, M., Jung, C.: Fully convolutional siamese fusion networks for object tracking. In: 2018 25Th IEEE international conference on image processing (ICIP). pp. 3718–3722. IEEE (2018)
4. Chen, X., Kang, B., Geng, W., Zhu, J., Liu, Y., Wang, D., Lu, H.: Sutrack: Towards simple and unified single object tracking. arXiv preprint arXiv:2412.19138 (2024)
5. Chen, X., Peng, H., Wang, D., Lu, H., Hu, H.: 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)
6. Chen, X., Sun, C., Xu, J., Peng, H., Wang, D., Lu, H., Ma, K.: Relo: Reinforcement learning to localize for visual object tracking. arXiv preprint arXiv:2605.07379 (2026)
7. Chen, X., Yan, B., Zhu, J., Lu, H., Ruan, X., Wang, D.: High-performance transformer tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(7), 8507–8523 (2022)
8. Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., Lu, H.: Transformer tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8126–8135 (2021)
9. Cui, Y., Jiang, C., Wang, L., Wu, G.: 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)
10. Cui, Y., Jiang, C., Wu, G., Wang, L.: Mixformer: End-to-end tracking with iterative mixed attention. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR. pp. 1–9 (2020)
12. Fan, H., Bai, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Harshit, Huang, M., Liu, J., et al.: Lasot: A high-quality large-scale single object tracking benchmark. International Journal of Computer Vision **129**(2), 439–461 (2021)
13. Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., Ling, H.: 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)
14. Feng, X., Hu, S., Li, X., Zhang, D., Wu, M., Zhang, J., Chen, X., Huang, K.: Atctrack: Aligning target-context cues with dynamic target states for robust vision-language tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19850–19861 (2025)
15. Feng, X., Li, X., Hu, S., Zhang, D., Wu, M., Zhang, J., Chen, X., Huang, K.: Memvlt: Vision-language tracking with adaptive memory-based prompts. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024)
16. Feng, X., Zhang, D., Hu, S., Li, X., Wu, M., Zhang, J., Chen, X., Huang, K.: Cstrack: Enhancing rgb-x tracking via compact spatiotemporal features. arXiv preprint arXiv:2505.19434 (2025)

17. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)
18. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In: International conference on learning representations (2018)
19. Guo, M., Tan, W., Ran, W., Jing, L., Zhang, Z.: Dreamtrack: Dreaming the future for multimodal visual object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7201–7210 (2025)
20. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR. pp. 16000–16009 (2022)
21. Hermann, K.L., Mobahi, H., Fel, T., Mozer, M.C.: On the foundations of shortcut learning. arXiv preprint arXiv:2310.16228 (2023)
22. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
23. Hu, S., Zhao, X., Huang, L., Huang, K.: Global instance tracking: Locating target more like humans. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(1), 576–592 (2023)
24. Huang, L., Zhao, X., Huang, K.: Got-10k: A large high-diversity benchmark for generic object tracking in the wild. IEEE transactions on pattern analysis and machine intelligence **43**(5), 1562–1577 (2019)
25. Huang, L., Zhao, X., Huang, K.: Globaltrack: A simple and strong baseline for long-term tracking. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 11037–11044 (2020)
26. Kang, B., Chen, X., Lai, S., Liu, Y., Liu, Y., Wang, D.: Exploring enhanced contextual information for video-level object tracking. In: Proceedings of the AAAI conference on artificial intelligence. vol. 39, pp. 4194–4202 (2025)
27. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7482–7491 (2018)
28. Kristan, M., Matas, J., Leonardis, A., Vojíř, T., Pflugfelder, R., Fernández, G., Nebehay, G., Porikli, F., Čehovin, L.: A novel performance evaluation methodology for single-target trackers. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(11), 2137–2155 (2016)
29. Li, B., Yan, J., Wu, W., Zhu, Z., Hu, X.: High performance visual tracking with siamese region proposal network. In: CVPR. pp. 8971–8980 (2018)
30. Li, X., Zhong, B., Liang, Q., Li, G., Mo, Z., Song, S.: Mambalct: Boosting tracking via long-term context state space model. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 4986–4994 (2025)
31. Liang, S., Bai, Y., Gong, Y., Wei, X.: Autoregressive sequential pretraining for visual tracking. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7254–7264 (2025)
32. Lin, L., Fan, H., Xu, Y., Ling, H.: Swintrack: A simple and strong baseline for transformer tracking. In: NeurIPS. pp. 16743–16754 (2022)
33. Lin, L., Fan, H., Zhang, Z., Huang, Y., Wang, Y., Xu, Y., Ling, H.: Loratv2: Enabling low-cost temporal modeling in one-stream trackers. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems
34. Lin, L., Fan, H., Zhang, Z., Wang, Y., Xu, Y., Ling, H.: Tracking meets lora: Faster training, larger model, stronger performance. In: ECCV. pp. 1–15 (2024)

35. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
36. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
37. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
38. Muller, M., Bibi, A., Giancola, S., Alsubaihi, S., Ghanem, B.: 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)
39. Peng, L., Gao, J., Liu, X., Li, W., Dong, S., Zhang, Z., Fan, H., Zhang, L.: Vasttrack: Vast category visual object tracking. arXiv preprint arXiv:2403.03493 (2024)
40. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 658–666 (2019)
41. Ross, A.S., Hughes, M.C., Doshi-Velez, F.: Right for the right reasons: Training differentiable models by constraining their explanations. arXiv preprint arXiv:1703.03717 (2017)
42. Shi, L., Zhong, B., Liang, Q., Li, N., Zhang, S., Li, X.: Explicit visual prompts for visual object tracking. In: AAAI. pp. 4838–4846 (2024)
43. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR 2011. pp. 1521–1528. IEEE (2011)
44. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS. pp. 5998–6008 (2017)
45. Wang, A., Liu, A., Zhang, R., Kleiman, A., Kim, L., Zhao, D., Shirai, I., Narayanan, A., Russakovsky, O.: Revise: A tool for measuring and mitigating bias in visual datasets. International Journal of Computer Vision **130**(7), 1790–1810 (2022)
46. Wang, X., Shu, X., Zhang, Z., Jiang, B., Wang, Y., Tian, Y., Wu, F.: 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 (CVPR). pp. 13763–13773 (June 2021)
47. Wei, X., Bai, Y., Zheng, Y., Shi, D., Gong, Y.: Autoregressive visual tracking. In: CVPR. pp. 9697–9706 (2023)
48. Wu, Y., Lim, J., Yang, M.H.: Object tracking benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(9), 1834–1848 (2015)
49. Xie, J., Zhong, B., Mo, Z., Zhang, S., Shi, L., Song, S., Ji, R.: Autoregressive queries for adaptive tracking with spatio-temporal transformers. In: CVPR. pp. 19300–19309 (2024)
50. Yan, B., Peng, H., Fu, J., Wang, D., Lu, H.: Learning spatio-temporal transformer for visual tracking. In: ICCV. pp. 10448–10457 (2021)
51. Ye, B., Chang, H., Ma, B., Shan, S., Chen, X.: Joint feature learning and relation modeling for tracking: A one-stream framework. In: ECCV. pp. 341–357. Springer (2022)
52. Ye, W., Zheng, G., Cao, X., Ma, Y., Zhang, A.: Spurious correlations in machine learning: A survey. arXiv e-prints pp. arXiv–2402 (2024)
53. Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., Finn, C.: Gradient surgery for multi-task learning. Advances in neural information processing systems **33**, 5824–5836 (2020)

54. Zhang, D., Hu, S., Feng, X., Li, X., Wu, M., Zhang, J., Huang, K.: Beyond accuracy: Tracking more like human via visual search. Advances in Neural Information Processing Systems **37**, 2629–2662 (2024)
55. Zhang, X., Tian, Y., Xie, L., Huang, W., Dai, Q., Ye, Q., Tian, Q.: Hivit: A simpler and more efficient design of hierarchical vision transformer. In: The Eleventh International Conference on Learning Representations (2023)
56. Zhang, Y., Gao, J., Wang, H., Ge, F., Luo, G., Hu, W., Zhang, Z.: Integrating diverse assignment strategies into detrs. arXiv preprint arXiv:2601.09247 (2026)
57. Zheng, Y., Liang, Q., Zhong, B., Zeng, S., Xue, Y., Li, N., Song, S.: Boosting self-supervised tracking with contextual prompts and noise learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 35197–35206 (2026)
58. Zheng, Y., Zhong, B., Liang, Q., Li, N., Song, S.: Decoupled spatio-temporal consistency learning for self-supervised tracking. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 10635–10643 (2025)
59. Zheng, Y., Zhong, B., Liang, Q., Mo, Z., Zhang, S., Li, X.: Odtrack: Online dense temporal token learning for visual tracking. In: AAAI. pp. 7588–7596 (2024)
60. Zheng, Y., Zhong, B., Liang, Q., Tang, Z., Ji, R., Li, X.: Leveraging local and global cues for visual tracking via parallel interaction network. IEEE Transactions on Circuits and Systems for Video Technology **33**(4), 1671–1683 (2022)