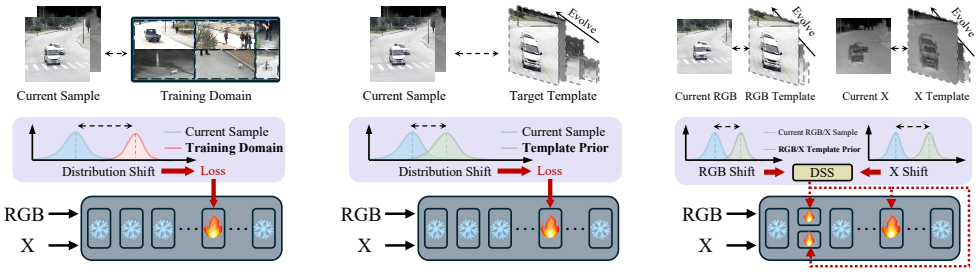
A STABLE SELF-EVOLVING TARGET TEMPLATE HELPS TEST-TIME ADAPTATION IN MULTIMODAL OBJECT TRACKING

Anonymous authors
Paper under double-blind review

ABSTRACT

RGB-X multimodal trackers have recently achieved remarkable progress by exploiting complementary information from heterogeneous modalities. However, existing methods often overlook performance degradation caused by distribution shifts between training and testing environments. Although recent test-time adaptation approaches have been explored for multimodal tracking, their adaptation objectives are often derived from test-time confidence estimates or source-domain statistics, which may become less reliable under significant distribution shifts. To address these challenges, we propose S^{2}TTA, a stable self-evolving target template-guided test-time adaptation framework. S^{2}TTA constructs a target-specific template prior and performs uncertainty-aware prior evolution to provide reliable test-time optimization objectives. Through a hierarchical adaptation space and discrepancy-guided space selection, S^{2}TTA adaptively selects appropriate adaptation spaces according to the distribution shifts of different modalities. Extensive experiments demonstrate that S^{2}TTA consistently achieves superior performance and significantly improves model robustness and generalization ability.



(a) Training-distribution Alignment (b) Template-guided Prior Alignment (c) S^{2}TTA

Figure 1: Illustration of different adaptation strategies. (a) Existing training-distribution alignment methods optimize the current sample toward the global training distribution, which consists of diverse objects and scenarios and may result in unreliable adaptation objectives. (b) We introduce template-guided prior alignment, which constructs a target-specific template prior and evolves it with temporal target information for stable adaptation. (c) Building upon the proposed template prior alignment, S^{2}TTA further introduces a hierarchical adaptation space with Discrepancy-guided Space Selection (DSS) to select appropriate spaces for different modality shifts.

1 INTRODUCTION

RGB-X multimodal object tracking aims to continuously localize targets by exploiting complementary information from heterogeneous sensors, with applications in real-world scenarios such as autonomous driving, intelligent surveillance, and robotic perception Altaf & Kim (2025); Huang & Hao (2021); Saha et al. (2022). In recent years, Transformer Vaswani et al. (2017) has been widely adopted to enhance representation learning and cross-modal interaction in multimodal tracking.

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.

However, existing multimodal trackers generally overlook the distribution shifts between training and testing environments Zaveri et al. (2025). In real-world scenarios, variations in illumination, weather, viewpoints, scenes, and target appearances can cause substantial gaps between training and testing distributions Li et al. (2019); Wang et al. (2025). Such shifts may degrade the reliability of learned representations, resulting in unstable feature responses and performance degradation during inference Mummadi et al. (2021). Moreover, the temporal nature of tracking further exacerbates error accumulation, causing progressive performance degradation over time Huang et al. (2026).

Test-time adaptation (TTA) provides an effective paradigm for mitigating performance degradation caused by distribution shifts by enabling pre-trained models to adapt during inference using only unlabeled test data Wang et al. (2020; 2022). Although recent TTA approaches have been extended to visual tracking, their adaptation objectives are still commonly derived from model-generated confidence estimates or training-domain statistics Qiao et al. (2026); Shao et al. (2025). Confidence-driven strategies heavily rely on the reliability of model predictions, whereas severe distribution shifts can cause inaccurate predictions and poorly calibrated confidence in the test environment, resulting in unreliable adaptation signals Lee et al. (2024; 2023). Furthermore, some methods attempt to alleviate distribution shifts by aligning test samples with the training data distribution Lin et al. (2023). However, the original training data is usually unavailable during test-time adaptation, and the training distribution typically aggregates diverse samples from various scenes, target categories, and environmental conditions Park et al. (2024); Fan et al. (2019); Muller et al. (2018). Directly aligning a specific test sample with such a broad and heterogeneous distribution may provide unreliable adaptation objectives, thereby limiting adaptation effectiveness Jung et al. (2023). Moreover, existing adaptation strategies often apply uniform updates across modalities, without explicitly considering modality-specific distribution shifts in multimodal tracking Li & Feng (2026).

To address these challenges, we propose S^2TTA, a Stable Self-evolving Target Template-guided Test-Time Adaptation framework for multimodal object tracking. S^2TTA consists of two key components: a stable self-evolving target template prior and a discrepancy-guided selection mechanism over the hierarchical adaptation space. Unlike existing TTA methods that align test samples with the global training distribution, as illustrated in Fig. 1a, S^2TTA constructs a target-specific template prior from the template image. However, a static template prior cannot adequately capture temporal variations throughout the tracking sequence. Therefore, we introduce uncertainty-aware prior evolution to incorporate temporal target information while suppressing unreliable updates from uncertain observations. As illustrated in Fig. 1b, the distribution shift is estimated by measuring the discrepancy between the current sample and the updated template prior, providing a target-specific and stable optimization objective for online adaptation. Furthermore, considering the modality-specific distribution shifts inherent in RGB-X tracking, we design a hierarchical adaptation space with a discrepancy-guided space selection strategy. Instead of uniformly adapting all modalities, S^2TTA selectively activates suitable adaptation spaces according to the shift severity of different modalities, enabling targeted optimization of modality-specific distribution deviations, as illustrated in Fig. 1c. By combining the evolving target-specific prior with selective multimodal adaptation, S^2TTA enables stable and targeted online adaptation while mitigating unreliable updates and error accumulation. We validate our approach on multiple RGB-X tracking benchmarks across different modalities and challenging distribution-shift scenarios, demonstrating consistent performance improvements, enhanced robustness, and stronger generalization ability.

Our main contributions are summarized as follows:

- We propose S^2TTA, a target-template-guided test-time adaptation framework for multimodal object tracking. To the best of our knowledge, S^2TTA is the first to exploit the inherent target template in tracking to construct a target-specific adaptation objective, establishing a new perspective for test-time adaptation in visual tracking.
- We develop a stable self-evolving template prior and a hierarchical adaptation space with discrepancy-guided space selection. The former maintains a stable adaptation objective under temporal variations, while the latter selects suitable adaptation spaces according to modality-specific distribution shifts, enabling targeted multimodal adaptation.
- Extensive experiments on multiple RGB-X multimodal tracking benchmarks demonstrate that the proposed method consistently improves robustness and generalization ability under diverse and challenging distribution-shift scenarios.

2 Related Work

Multimodal Object Tracking Multimodal object tracking exploits complementary information from heterogeneous sensors to alleviate the limitations of RGB tracking under challenging conditions, such as low illumination, occlusion, and complex backgrounds Zhang et al. (2024a); Li et al. (2018; 2016). By integrating diverse modality cues, multimodal trackers have achieved promising results in various applications, including autonomous driving, robotic perception, and intelligent surveillance Zhang et al. (2024b); Saha et al. (2022); Chiu et al. (2021); Huang & Hao (2021).

Recently, Transformer Vaswani et al. (2017) has become a dominant architecture in multimodal tracking due to its global modeling and flexible interaction capabilities. Representative methods have explored visual prompt tuning Zhu et al. (2023), spatio-temporal representation learning Feng et al. (2025); Lu et al. (2026), efficient two-stream architectures Su et al. (2026), language-enhanced tracking Li et al. (2026a), and efficient cross-modal interaction with hybrid or mixture-of-experts architectures Li et al. (2026b); Tan et al. (2025a). These advances have substantially improved multimodal representation, cross-modal interaction, and tracking efficiency.

Despite this progress, existing multimodal trackers are generally developed under the assumption of similar training and testing distributions. In practice, variations in environments, sensors, viewpoints, and target appearances can introduce distribution shifts, leading to unstable representations and degraded tracking performance Wang et al. (2020). Therefore, adapting multimodal trackers to unknown test environments without additional annotations remains an important challenge.

Test-time Adaptation Test-time adaptation (TTA) has recently emerged as an effective paradigm for improving model robustness under distribution shifts Liang et al. (2025); Lim et al. (2023); Lee et al. (2023). Its goal is to enable pre-trained models to dynamically adapt to changing test environments during inference using only unlabeled test data, without requiring additional annotations or retraining. Recent studies have demonstrated the effectiveness of TTA in various scenarios, including image classification, semantic segmentation, and vision-language models Wang et al. (2020); Niu et al. (2022); Shu et al. (2022); Zhang et al. (2022); Karmanov et al. (2024).

Recently, several studies have explored extending TTA to temporal and multimodal tasks. For example, FSTTA Gao et al. (2023) introduced a fast-slow adaptation framework to balance rapid adaptation and long-term stability, while TCEI Guo et al. (2026) leveraged short-term memory and long-term experience for multi-object tracking adaptation. PURA Shao et al. (2025) further investigated RGB-T tracking adaptation with a parameter update-recovery mechanism to alleviate long-term drift. Beyond tracking, IDEA Hu et al. (2026) explored TTA in vision-and-language navigation by accumulating historical knowledge for stable adaptation under non-stationary environments.

However, these approaches fundamentally rely on model predictions or directly align test samples with source-domain statistics to construct adaptation signals. Under severe distribution shifts, predictions generated by pre-trained models may become unreliable, while source-domain alignment may introduce inaccurate objectives due to the broad and heterogeneous training distribution. Consequently, existing TTA strategies may derive unreliable optimization directions and accumulate adaptation errors, limiting their robustness in complex multimodal tracking scenarios.

3 Method

3.1 Overview

We propose S^{2}TTA, a target-template-guided test-time adaptation framework for RGB-X multimodal object tracking. It can be integrated with existing trackers to perform online adaptation under unknown distribution shifts. We first describe the inference pipeline of a typical dual-stream RGB-X multimodal tracker, as illustrated in Fig. 2. In single-stream architectures, cross-modal fusion is usually performed before encoder processing, and the corresponding pipeline is not elaborated here. Given the RGB image $\mathbf{I}_r^t$ and the auxiliary-modal image $\mathbf{I}_x^t$ at frame t, the two inputs are first projected into feature embeddings $\mathbf{Z}_r^t$ and $\mathbf{Z}_x^t$ through the patch embedding layer. For the template frame, the superscript t in this and the subsequent formulations is replaced by T. The modality-specific embeddings are subsequently processed by the pre-trained encoder $\mathcal{B}_\Theta$, producing encoded representations $\mathbf{Z}_{r,e}^t$ and $\mathbf{Z}_{x,e}^t$. These representations are then integrated by the cross-modal fusion

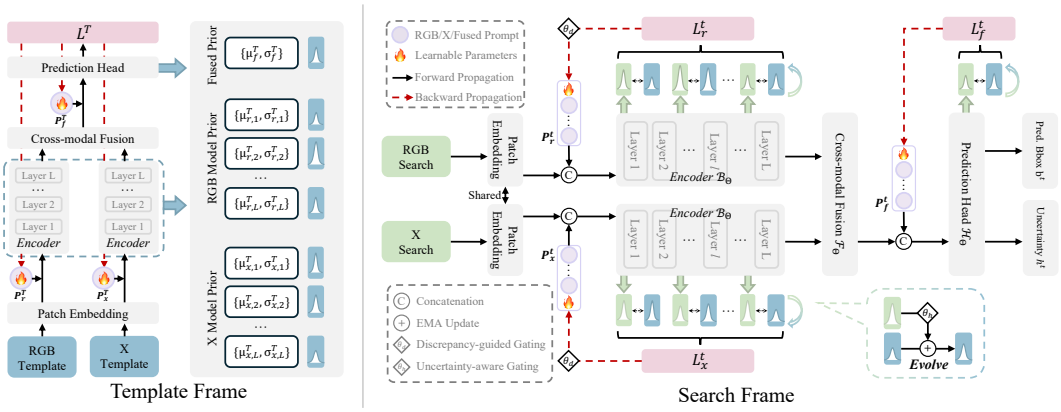


Figure 2: Overview of the proposed S^{2}TTA framework. Given the template frame, S^{2}TTA optimizes modality-specific prompts $\mathbf{P}_r$, $\mathbf{P}_x$ and fused prompt $\mathbf{P}_f$ using the pre-trained tracker and training objective, and constructs a target-specific multi-level template prior from optimized statistics. For each search frame, S^{2}TTA estimates modality distribution discrepancy and employs discrepancy-guided space selection strategy to select adaptation spaces for prompt updating. The target template prior is further updated with uncertainty-aware statistics for robust online test-time adaptation.

module $\mathcal{F}_\Theta$ to obtain the fused representation $\mathbf{Z}_f^t$. Finally, the prediction head $\mathcal{H}_\Theta$ produces the target response map $\mathbf{S}^t$, which consists of N response regions $\mathbf{S}_i^t$ and their corresponding weights $\mathbf{W}_i^t$, where $i \in \{1, \ldots, N\}$. The $\mathbf{S}^t$ is decoded to obtain the predicted target bounding box $\mathbf{b}^t$.

Our goal is to adapt the tracker to unknown test-time distribution shifts through lightweight learnable prompts while keeping the original pre-trained parameters Θ fixed. To enable efficient online adaptation, we first define a hierarchical adaptation space with learnable prompts across different modality branches, as introduced in Sec. 3.2. Test-time adaptation aims to mitigate performance degradation caused by distribution shifts, but obtaining an appropriate adaptation objective remains challenging in online tracking, as the original training data are typically unavailable during inference and retained source-domain statistics may provide only a coarse reference that does not accurately reflect the current target. RGB-X object tracking, however, provides an additional source of target-specific information: the annotated template frame. In Sec. 3.3, we process the template frame with the pre-trained tracker to establish a target-conditioned representation prior, which is specialized to the designated target. We therefore use the representation prior as a sequence-specific reference rather than attempting to recover the unavailable global training distribution. Under changing test conditions, deviations of subsequent target representations from this reference provide an observable proxy for the impact of distribution shifts on the current target representation. To enable reliable adaptation throughout the dynamic tracking process, Sec. 3.4 further introduces selective online adaptation with discrepancy-guided space selection and uncertainty-aware prior evolution.

3.2 Hierarchical Adaptation Space

Following the inference pipeline described above, we define a hierarchical adaptation space with lightweight learnable prompts for online optimization. Specifically, given the modality-specific feature embeddings $\mathbf{Z}_r^t$ and $\mathbf{Z}_x^t$ extracted by the pre-trained tracker, we introduce modality-specific prompts $\mathbf{P}_r^t$ and $\mathbf{P}_x^t$ before the corresponding encoders. After obtaining the fused representation through the fusion module, we further introduce a fused prompt $\mathbf{P}_f^t$ before the prediction head. Therefore, the overall inference pipeline with hierarchical prompts can be formulated as:

$$\mathbf{b}^t = \mathcal{H}_\Theta\left(\left[\mathbf{P}_f^t, \mathcal{F}_\Theta\left(\hat{\mathbf{Z}}_{r,e}^t, \hat{\mathbf{Z}}_{x,e}^t\right)\right]\right) = \mathcal{H}_\Theta\left(\left[\mathbf{P}_f^t, \mathcal{F}_\Theta\left(\mathcal{B}_\Theta\left(\left[\mathbf{P}_r^t, \mathbf{Z}_r^t\right]\right), \mathcal{B}_\Theta\left(\left[\mathbf{P}_x^t, \mathbf{Z}_x^t\right]\right)\right)\right]\right), \quad (1)$$

where the modality-specific prompts and the fused prompt constitute two adaptation levels with different influences on the tracking representation. The modality-specific prompts are inserted at an earlier stage and affect all feature encoding and subsequent cross-modal interactions. They therefore offer stronger adaptation capacity to severe modality shifts, but their longer update paths may

induce larger representation variations. In contrast, the fused prompt is introduced at a later stage, resulting in a shorter update path and a gentler adaptation of the model representations. Meanwhile, because it operates on the shared representation, it can influence the jointly encoded content of both modalities. This hierarchy provides the structural basis for selectively adapting different adaptation spaces according to the current model state. All prompts are parameterized in a unified form:

$$\mathbf{P}_b^t = \mathbf{P}_b^T + \Delta\mathbf{P}_b^t, \qquad b \in \{r, x, f\}, \tag{2}$$

where $\mathbf{P}_b^T$ denotes the template-conditioned prompt initialization, which remains fixed during online adaptation. The residual $\Delta\mathbf{P}b^t$ denotes the learnable prompt adjustment optimized by the template-prior alignment loss introduced in Sec. 3.4, and is initialized to zero before online adaptation.

3.3 Build template prior

After defining the hierarchical adaptation space with modality-specific and fusion prompts, the next challenge is to establish an appropriate adaptation reference for each test sequence. To this end, we exploit the annotated template frame to initialize the prompts and construct a target-conditioned representation prior, which provides a sequence-specific reference for subsequent online adaptation. A straightforward solution is to directly extract the template representation using the pre-trained tracker and use it as the reference for subsequent frames. However, the template frame itself belongs to the test domain and may exhibit distribution shifts from the training data, making the representation directly extracted by the pre-trained tracker an unreliable reference for subsequent adaptation. To obtain a template-conditioned representation adapted to the current test sequence, we leverage the template annotation to optimize the prompts within the introduced adaptation space. The resulting features are better aligned with the target localization objective, based on which we further construct the template prior by extracting fine-grained distribution statistics across multiple representation layers to provide a target-specific reference for subsequent adaptation.

Template-Conditioned Prompt Initialization. We first construct an augmentation pool

$$\mathcal{A}_0 = \{(\mathbf{I}_a^T, \mathbf{Y}_a^T)\}_{a=1}^M, \tag{3}$$

where $\mathbf{I}_a^T$ and $\mathbf{Y}_a^T$ denote the augmented template image and its corresponding target bounding box, respectively. The augmentation pool contains the original template image and $M-1$ additional augmented views, which prevents $\mathbf{P}_b^T$ from overfitting to a single template observation. The detailed augmentation strategies and corresponding analysis are provided in Appendix A.3. For each augmented view, let $\mathbf{b}_a^T$ denote the predicted bbox. We optimize the template-conditioned prompts $\mathbf{P}_b^T$ using the original tracking objective, which is formulated as:

$$\mathcal{L}^T = \mathcal{L}_{\text{cls}}(\mathbf{S}_a^T, \mathbf{Y}_a^T) + \lambda_{\text{iou}}\mathcal{L}_{\text{iou}}(\mathbf{b}_a^T, \mathbf{Y}_a^T) + \lambda_{\text{L1}}\mathcal{L}_{\text{L1}}(\mathbf{b}_a^T, \mathbf{Y}_a^T). \tag{4}$$

Multi-Level Distribution Template Prior. Given the initialized template-conditioned prompts $\mathbf{P}_b^T$, we process all augmented template images through the corresponding modality encoders $\mathcal{B}_\Theta$ and extract hierarchical representations from the encoder layers and the final fused representation before the prediction head. The statistics averaged over the augmentation pool constitute the multi-level template prior, denoted as $\{(\boldsymbol{\mu}_{b,l}^T, \boldsymbol{\sigma}_{b,l}^T)\}_{l=1}^{L_b}$. Since different representation levels contribute unequally to target prediction, we assign a weight w_l to each level and aggregate the weighted statistics to obtain the final template prior $\{\boldsymbol{\mu}_b^T, \boldsymbol{\sigma}_b^T\}$. The detailed weighting strategy and the comparison analysis of different weighting schemes are provided in Appendix A.1.

3.4 Selective Online Adaptation

Based on the observation described in Sec. 3.1, we use the discrepancy between current-frame representation statistics and the template prior as a proxy for target-specific distribution shifts. Accordingly, S^{2}TTA aligns the current representation with the template prior to guide online adaptation, while selecting appropriate adaptation spaces according to modality-specific distribution discrepancies. Meanwhile, the uncertainty of model predictions is leveraged to determine whether the template prior should evolve over time, preventing unreliable updates under unstable observations.

Distribution Discrepancy and Template Prior Alignment. For each frame t, we extract the same multi-level representation statistics used for prior construction. We define the discrepancy between the current representation and the corresponding template prior as

$$D_b^t = \sum_{l=1}^{L_b} w_l \left(\left\| \boldsymbol{\mu}_{b,l}^t - \boldsymbol{\mu}_{b,l}^T \right\|_2 + \alpha_\sigma \left\| \boldsymbol{\sigma}_{b,l}^t - \boldsymbol{\sigma}_{b,l}^T \right\|_2 \right), \tag{5}$$

where $\boldsymbol{\mu}_{b,l}^t$ and $\boldsymbol{\sigma}_{b,l}^t$ denote the current-frame statistics at level l, and α_σ controls the relative contribution of the standard-deviation term. Based on the above discrepancy, we directly formulate the template prior alignment objectives as $\mathcal{L}_b^t = D_b^t$. This yields a target-specific alignment objective without requiring access to the original training data for online adaptation.

Prediction-Uncertainty Assessment. We determine whether to update the template prior based on prediction uncertainty estimated by the entropy of the current-frame response map. A lower entropy indicates a more concentrated response distribution and hence a more stable prediction. Specifically, following the response map $\mathbf{S}^t$ described in Sec. 3.1, we normalize the weighted responses as $q_i^t = \mathrm{Softmax}_i\left(\log(\mathbf{W}_i^t \mathbf{S}_i^t)\right)$, and compute the uncertainty as $h^t = -\sum_i q_i^t \log q_i^t$.

Discrepancy-Guided Space Selection and Uncertainty-Aware Prior Evolution. The D_b^t quantify the severity of distribution shifts across different modalities, while the entropy h_t reflects the uncertainty of model predictions. As discussed in Sec. 3.2, prompts at different branches have complementary responsibilities and adaptation strengths. We therefore use the D_b^t to select the appropriate adaptation space, and use the h_t to determine whether to update the template prior:

$$\mathcal{C}_p^t = \{f\} \cup \left\{ c \in \{r, x\} \mid D_c^t > \theta_d \right\}, \quad \mathcal{C}_a^t = \mathbb{I}\left(h^t \le \theta_h\right), \tag{6}$$

where $\mathcal{C}_p^t$ denotes the selected adaptation space for the current frame, and $\mathcal{C}_a^t$ indicates whether the template prior is updated. As discussed in Sec. 3.2, the fused prompt is always activated to adjust the shared multimodal representation, while modality-specific prompts are additionally selected when their corresponding discrepancies exceed θ_d. This design enables adaptive combinations of fusion-level and modality-level adaptation for different shift severities. Meanwhile, when the uncertainty h^t is below the threshold θ_h, indicating stable model predictions, the current distribution statistics are incorporated into the template prior through exponential moving averaging. Otherwise, the template prior update is skipped for the current frame to prevent prior drift and error accumulation. The corresponding prior evolution formulations is given as follows:

$$\left(\boldsymbol{\mu}_b^T, \boldsymbol{\sigma}_b^T\right) \leftarrow \begin{cases} (1-m)\left(\boldsymbol{\mu}_b^T, \boldsymbol{\sigma}_b^T\right) + m(\boldsymbol{\mu}_b^t, \boldsymbol{\sigma}_b^t), & \mathcal{C}_a^t = 1, \\ \left(\boldsymbol{\mu}_b^T, \boldsymbol{\sigma}_b^T\right), & \mathcal{C}_a^t = 0. \end{cases} \tag{7}$$

where m is a momentum coefficient that controls the update intensity of the template prior. In this way, by leveraging the template prior to provide a optimization objective and using discrepancy-guided and uncertainty-aware mechanisms to select the hierarchical adaptation space and update the template prior, our method achieves robust adaptation to unseen test environments.

4 Experiments

4.1 Experimental Setup

Datasets and Metrics. To evaluate the effectiveness of S^{2}TTA across different multimodal settings, we conduct experiments on three types of RGB-X tracking scenarios with different auxiliary modalities, including RGB-T, RGB-D, and RGB-E. Specifically, we evaluate our method on widely used benchmarks in previous multimodal tracking studies, including RGBT234 Li et al. (2019) and LasHeR Li et al. (2021) with thermal infrared modalities, DepthTrack Yan et al. (2021) and VOT-RGBD22 Kristan et al. (2022) with depth modalities, and VisEvent Wang et al. (2023) with event modalities. Following the common evaluation protocols in existing multimodal tracking studies, we adopt the standard metrics for each benchmark to ensure fair comparisons.

Table 1: Comparisons with state-of-the-art methods on RGB-T, RGB-D, and RGB-E benchmarks, including LasHeR, RGBT234, DepthTrack, VOT-RGBD2022, and VisEvent. The best results are highlighted with bold fonts. The two baseline methods are indicated by gray backgrounds.

Method	Publication	RGB-T				RGB-D						RGB-E	
		LasHeR		RGBT234		DepthTrack			VOT-RGBD22			VisEvent	
		PR (↑)	AUC (↑)	MPR (↑)	MSR (↑)	Pr (↑)	Re (↑)	F-score (↑)	EAO (↑)	Acc. (↑)	Rob. (↑)	SR (↑)	PR (↑)
OSTrack Ye et al. (2022)	ECCV'22	53.0	42.2	75.5	56.9	56.9	58.2	55.7	66.6	80.8	81.4	53.4	69.5
ViPT Zhu et al. (2023)	CVPR'23	65.1	52.5	83.5	61.7	59.2	59.6	59.4	72.1	81.5	87.1	59.2	75.8
Un-Track Wu et al. (2024)	CVPR'24	66.7	53.6	83.7	61.8	61.3	61.0	61.2	72.1	81.5	87.1	58.9	75.5
SDSTrack Hou et al. (2024)	CVPR'24	66.5	53.1	84.8	62.5	61.9	60.9	61.4	72.8	81.2	88.3	59.7	76.7
OneTracker Hong et al. (2024)	CVPR'24	67.2	53.8	85.7	64.2	60.7	60.4	60.9	72.7	81.9	87.2	60.8	76.7
TBSI Hui et al. (2023)	CVPR'24	70.5	56.3	86.4	64.3	–	–	–	–	–	–	–	–
BAT Cao et al. (2024)	AAAI'24	70.2	56.3	86.8	64.1	–	–	–	–	–	–	–	–
GMMT Tang et al. (2024)	AAAI'24	70.7	56.6	87.9	64.7	–	–	–	–	–	–	–	–
SEATrack Su et al. (2026)	CVPR'26	71.6	57.3	87.8	63.9	62.9	63.5	63.2	73.6	82.1	88.4	77.1	60.3
XTrack Tan et al. (2025b)	ICCV'25	73.1	58.7	87.8	65.4	65.4	64.3	64.8	74.0	82.8	88.9	63.3	80.5
STTrack Hu et al. (2025)	AAAI'25	76.0	60.3	89.8	66.7	63.3	63.4	63.2	77.6	82.5	93.7	–	–
CSTrack(Reproduce) Feng et al. (2025)	ICML'25	74.8	60.0	93.5	70.2	65.1	66.3	65.7	77.3	83.1	92.8	65.0	82.3
CSTrack Feng et al. (2025)	ICML'25	75.6	60.8	94.0	70.9	65.2	66.4	65.8	77.4	83.3	92.9	65.2	82.4
SEATrack + S^{2}TTA	Ours	73.0	58.2	88.5	66.1	64.1	64.1	64.2	74.2	82.6	88.9	**77.7**	**60.7**
CSTrack(Reproduce) + S^{2}TTA	Ours	**76.0**	**60.8**	**94.8**	**72.8**	**66.2**	**66.8**	**66.5**	**77.9**	**83.5**	**93.2**	65.5	82.6

Table 2: Cross-domain evaluation of S^{2}TTA under severe distribution shifts. Here, ✓ denotes the baseline tracker equipped with S^{2}TTA, while × denotes the original baseline tracker.

Train-domain	S^{2}TTA	Test-domain											
		RGB-T				RGB-D						RGB-E	
		LasHeR		RGBT234		DepthTrack			VOT-RGBD22			VisEvent	
		PR	AUC	MPR	MSR	Pr	Re	F-score	EAO	Acc	Rob	SR	PR
RGB-T (LasHeR)	×	–	–	–	–	52.5	50.0	51.7	63.8	75.9	80.1	43.4	25.9
	✓	–	–	–	–	**53.5**	**51.1**	**52.7**	**65.0**	**77.1**	**81.2**	**53.5**	**35.6**
RGB-E (VisEvent)	×	53.9	44.1	76.6	56.8	47.4	48.0	47.7	66.5	78.2	82.9	–	–
	✓	**56.2**	**45.6**	**79.3**	**59.5**	**48.5**	**49.1**	**48.7**	**67.7**	**79.4**	**84.1**	–	–
RGB-D (DepthTrack)	×	53.6	41.8	75.5	56.2	–	–	–	–	–	–	48.7	32.1
	✓	**57.0**	**43.5**	**77.7**	**58.2**	–	–	–	–	–	–	**54.9**	**37.9**

Implementation Details. We adopt two recent and representative state-of-the-art trackers, CSTrack Feng et al. (2025) and SEATrack Su et al. (2026), as our baselines. CSTrack represents a typical single-stream architecture, where multiple modalities are jointly modeled within a unified representation learning framework and trained across diverse multimodal data. In contrast, SEATrack adopts a typical dual-stream architecture and trains modality-specific models for different multimodal tracking settings. Together, these two baselines span representative single-stream and dual-stream multimodal tracking architectures, allowing us to evaluate the general applicability of S^{2}TTA across different model designs. The thresholds in the selective online adaptation are set to $\theta_d = 0.7$ and $\theta_h = 2.2$, respectively. The momentum coefficient for stable prior evolution is set to 0.03. We set $L_r = L_x = 4$ for modality-specific representations and $L_f = 1$ for the fused representation. All experiments are conducted on a single NVIDIA RTX 4090 GPU.

4.2 Main Results

Comparison with State-of-the-art Methods. As shown in Table 1, S^{2}TTA achieves state-of-the-art performance across all evaluated RGB-X multimodal tracking benchmarks, consistently outperforming existing methods on diverse modalities and scenarios. Recently proposed MDTrack Wang et al. (2026) is not included in the comparison since its source code is unavailable for reproduction and further evaluation within our framework. Furthermore, compared with the original baseline trackers, S^{2}TTA consistently brings additional improvements by introducing test-time adaptation. Specifically, S^{2}TTA improves the performance on LasHeR, RGBT234, and DepthTrack by 1.2%, 2.6%, and 1.1%, respectively, demonstrating significant gains under challenging distribution shifts. These results validate that the proposed target-template-guided adaptation framework can effectively enhance existing RGB-X trackers through stable prior construction and selective online adaptation.

Cross-Domain Evaluation Under Severe Distribution Shifts. To further evaluate the robustness of S^{2}TTA under more challenging distribution shifts, we conduct cross-domain experiments by training the tracker on one modality domain and directly testing it on unseen domains with different modalities. As shown in Table 2, when the training and testing domains are inconsistent,

Table 3: Comparison with representative TTA methods and ablation studies of different components in S^{2}TTA. The gray background denotes the complete S^{2}TTA framework.

Method	Template Prior	Adaptation Space	RGB-T				RGB-D						RGB-E	
			LasHeR		RGBT234		DepthTrack			VOT-RGBD22			VisEvent	
			PR	AUC	MPR	MSR	Pr	Re	F-score	EAO	Acc	Rob	SR	PR
Baseline			74.8	60.0	93.5	70.2	65.1	66.3	65.7	77.3	83.1	92.8	65.0	82.3
Tent Wang et al. (2020)			74.5	59.5	94.1	71.0	64.8	65.9	65.5	76.9	82.6	92.4	64.7	81.6
EATA Niu et al. (2022)			74.6	59.8	94.2	71.0	64.7	65.7	65.2	77.0	82.8	92.4	64.8	81.8
PURA Shao et al. (2025)			75.0	60.2	93.8	70.6	65.3	66.2	65.9	77.1	82.9	92.6	65.0	82.2
IDEA Hu et al. (2026)			74.8	60.0	93.3	70.2	65.4	63.7	62.9	77.4	83.3	92.9	65.1	82.3
S^{2}TTA (Ours)	✓	✓	**76.0**	**60.8**	**94.8**	**72.8**	**66.2**	**66.8**	**66.5**	**77.9**	**83.5**	**93.2**	**65.5**	**82.6**
S^{2}TTA (Ours)	✓	–	75.5	60.4	94.4	71.3	65.7	66.5	66.2	77.6	83.3	93.0	65.2	82.4
S^{2}TTA (Ours)	–	✓	75.0	60.1	93.9	70.9	65.0	66.2	65.7	77.5	83.0	92.8	65.0	82.3

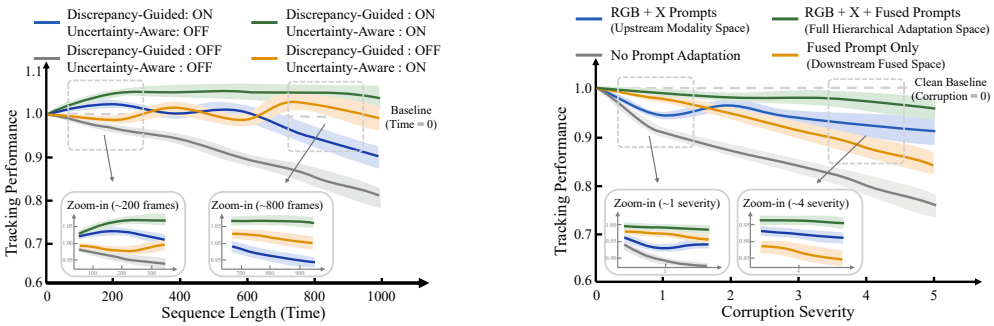
the baseline(SEATrack) tracker suffers from significant performance degradation due to the large modality and distribution gaps. In contrast, equipping the baseline tracker with S^{2}TTA consistently alleviates the performance drop across different cross-domain settings, achieving substantial improvements over the original tracker. In particular, S^{2}TTA delivers an improvement of approximately 10.1% on the VisEvent dataset. The relatively smaller gains observed on the other datasets may be attributed to differences in baseline performance, modality characteristics, and training-set scale. These results demonstrate that the proposed strategy can effectively provide reliable adaptation objectives and improve model robustness under severe unknown distribution shifts. It is worth noting that training-distribution Alignment based TTA methods are not included in this evaluation, since directly aligning cross-domain test samples with the original source distribution may introduce misleading adaptation objectives under such severe domain discrepancies.

Comparison with Other TTA Methods. To evaluate the effectiveness of S^{2}TTA against existing test-time adaptation approaches, we compare it with four representative TTA methods, including the widely adopted Tent Wang et al. (2020) and EATA Niu et al. (2022) for classification, PURA Shao et al. (2025), a recently proposed TTA method for multimodal object tracking, and IDEA Hu et al. (2026), a recent source-domain alignment based TTA method for vision-language navigation. Due to space constraints, the efficiency comparison with these TTA methods is provided in Appendix A.2. As shown in Table 3, existing TTA methods exhibit inconsistent adaptation performance across multimodal tracking benchmarks. Their improvements are generally limited, and some methods even suffer from performance degradation on certain datasets, indicating that directly transferring existing TTA strategies to multimodal tracking may lead to unreliable adaptation objectives. In contrast, S^{2}TTA consistently achieves improvements across all evaluated benchmarks with larger gains over the corresponding baselines. These results demonstrate the robustness and effectiveness of S^{2}TTA for handling complex distribution shifts in multimodal object tracking.

4.3 Ablation Studies

Component Ablation. To investigate the effectiveness of each component in S^{2}TTA, we conduct ablation studies by comparing the complete framework with two variants. Specifically, the first variant only adopts the stable self-evolving target template prior to construct the adaptation objective, without using the hierarchical adaptation space and discrepancy-guided space selection strategy. The second variant only employs the hierarchical adaptation space with discrepancy-guided space selection strategy, while replacing the template-guided optimization objective with the conventional training-distribution alignment objective. The results are reported in Table 3.

As shown in Table 3, the template-prior alignment objective alone consistently improves the baseline, demonstrating that a reliable target-specific optimization objective is fundamental to effective test-time adaptation. In contrast, introducing the hierarchical adaptation space without the proposed template-guided objective leads to less stable performance, indicating that increased adaptation flexibility alone does not necessarily guarantee reliable improvements. When both components are jointly incorporated, S^{2}TTA achieves the best performance, suggesting that the template prior provides a reliable optimization direction, while the hierarchical adaptation space offers greater capacity to exploit this objective through modality-specific and selectively expanded adaptation.



(a) Long-term tracking performance analysis of selective online adaptation.

(b) Robustness analysis of hierarchical adaptation spaces under modality corruption.

Figure 3: Analysis of the key adaptation mechanisms in S^{2}TTA, examining their long-term stability and behavior under different distribution shifts simulated by varying corruption levels.

Long-Term Analysis of Selective Online Adaptation. We analyze the long-term effect of discrepancy-guided space selection and uncertainty-aware prior evolution by evaluating their different combinations over increasing sequence lengths. As shown in Fig. 3a, jointly enabling both mechanisms yields the most stable performance over long sequences. Removing uncertainty-aware prior evolution leads to progressive degradation, indicating that a static target prior cannot adequately accommodate temporal variations in target representations, resulting in increasing prior mismatch and error accumulation. By contrast, using uncertainty-aware prior evolution alone preserves long-term stability but limits adaptation capacity, as the adaptation space cannot be appropriately selected according to the distribution discrepancies of different branches. Disabling both mechanisms results in the largest performance drop. These results confirm their complementary roles: discrepancy-guided space selection improves adaptation flexibility across different modality shifts, while uncertainty-aware prior evolution suppresses accumulated errors and maintains a stable optimization reference.

Analysis of the Hierarchical Adaptation Space. We analyze the performance of different prompt adaptation spaces under varying degrees of distribution shift. Specifically, we introduce Gaussian noise with increasing severity to the input samples to simulate progressively stronger distribution shifts. As shown in Fig.3b, jointly adapting the modality-specific and fused prompts consistently achieves the best overall robustness, with only mild degradation as the shift becomes more severe. When only the upstream modality-specific prompts are adapted, performance fluctuates more noticeably under mild shifts but becomes relatively stable as the shift severity increases. This behavior suggests that upstream adaptation provides stronger capacity for correcting severe modality-specific shifts, while its aggressive adjustment may introduce unnecessary representation perturbations under mild shifts. In contrast, adapting only the downstream fused prompt performs well under mild shifts but degrades rapidly as the shift becomes severe, indicating that its more conservative adjustment is insufficient to compensate for strong modality-level representation changes. These observations are consistent with the different roles of the two adaptation levels discussed in Sec.3.2, further supporting the motivation for the proposed hierarchical adaptation space.

5 Conclusion

In this paper, we address the performance degradation of multimodal object tracking under distribution shifts and propose S^{2}TTA, a stable self-evolving target template-guided test-time adaptation framework. By exploiting the inherent target template in visual tracking, S^{2}TTA constructs a target-specific template prior and performs uncertainty-aware prior evolution to provide a reliable optimization objective for online adaptation under unknown test environments. Meanwhile, the hierarchical adaptation space with discrepancy-guided space selection enables targeted updates according to modality-specific distribution shifts. Extensive experiments on multiple benchmarks demonstrate that S^{2}TTA consistently improves robustness and generalization. Additional discussions on the limitations of S^{2}TTA and possible future research directions are included in Appendix B.

AI USE STATEMENT

Generative AI tools were used solely for language polishing and improving the clarity and readability of the manuscript. All AI-assisted revisions were carefully reviewed and modified by the authors, who take full responsibility for the final content of this work.

REFERENCES

Muhammad Adeel Altaf and Min Young Kim. Multiple object detection and tracking in autonomous vehicles: A survey on enhanced affinity computation and its multimodal applications. *ICT Express*, 2025.

Bing Cao, Junliang Guo, Pengfei Zhu, and Qinghua Hu. Bi-directional adapter for multimodal tracking. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 927–935, 2024.

Hsu-kuang Chiu, Jie Li, Rareş Ambruş, and Jeannette Bohg. Probabilistic 3d multi-modal, multi-object tracking for autonomous driving. In *2021 IEEE international conference on robotics and automation (ICRA)*, pp. 14227–14233. IEEE, 2021.

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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5369–5378. IEEE, 2019.

Xiaokun Feng, Dailing Zhang, Shiyu Hu, Xuchen Li, Meiqi Wu, Jing Zhang, Xiaotang Chen, and Kaiqi Huang. Cstrack: Enhancing rgb-x tracking via compact spatiotemporal features. *arXiv preprint arXiv:2505.19434*, 2025.

Junyu Gao, Xuan Yao, and Changsheng Xu. Fast-slow test-time adaptation for online vision-and-language navigation. *arXiv preprint arXiv:2311.13209*, 2023.

Wen Guo, Pengfei Zhao, Zongmeng Wang, Yufan Hu, and Junyu Gao. Dual-level adaptation for multi-object tracking: Building test-time calibration from experience and intuition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28190–28200, 2026.

Lingyi Hong, Shilin Yan, Renrui Zhang, Wanyun Li, Xinyu Zhou, Pinxue Guo, Kaixun Jiang, Yiting Chen, Jinglun Li, Zhaoyu Chen, et al. Onetracker: Unifying visual object tracking with foundation models and efficient tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19079–19091. IEEE, 2024.

Xiaojun Hou, Jiazheng Xing, Yijie Qian, Yaowei Guo, Shuo Xin, Junhao Chen, Kai Tang, Mengmeng Wang, Zhengkai Jiang, Liang Liu, et al. Sdstrack: Self-distillation symmetric adapter learning for multi-modal visual object tracking. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26541–26551. IEEE, 2024.

Xiantao Hu, Ying Tai, Xu Zhao, Chen Zhao, Zhenyu Zhang, Jun Li, Bineng Zhong, and Jian Yang. Exploiting multimodal spatial-temporal patterns for video object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3581–3589, 2025.

Zixuan Hu, Xuantuo Huang, Yancheng Li, Yichun Hu, Shengyong Xu, and Ling-Yu Duan. Turning adaptation into assets: Cross-domain bridging for online vision-language navigation. *arXiv preprint arXiv:2605.23257*, 2026.

Kemiao Huang and Qi Hao. Joint multi-object detection and tracking with camera-lidar fusion for autonomous driving. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6983–6989. IEEE, 2021.

Yuqing Huang, Liting Lin, Weijun Zhuang, Zhenyu He, and Xin Li. Drift-resilient temporal priors for visual tracking. *arXiv preprint arXiv:2604.02654*, 2026.

Tianrui Hui, Zizheng Xun, Fengguang Peng, Junshi Huang, Xiaoming Wei, Xiaolin Wei, Jiao Dai, Jizhong Han, and Si Liu. Bridging search region interaction with template for rgb-t tracking. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13630–13639. IEEE, 2023.

Sanghun Jung, Jungsoo Lee, Nanhee Kim, Amirreza Shaban, Byron Boots, and Jaegul Choo. Cafa: Class-aware feature alignment for test-time adaptation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19014–19025. IEEE, 2023.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14162–14171, 2024.

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.

Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In *International Conference on Learning Representations*, volume 2024, pp. 46977–47002, 2024.

Jungsoo Lee, Debasmit Das, Jaegul Choo, and Sungha Choi. Towards open-set test-time adaptation utilizing the wisdom of crowds in entropy minimization. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16334–16334. IEEE, 2023.

Chenglong Li, Hui Cheng, Shiyi Hu, Xiaobai Liu, Jin Tang, and Liang Lin. Learning collaborative sparse representation for grayscale-thermal tracking. *IEEE Transactions on Image Processing*, 25 (12):5743–5756, 2016.

Chenglong Li, Xiaohao Wu, Nan Zhao, Xiaochun Cao, and Jin Tang. Fusing two-stream convolutional neural networks for rgb-t object tracking. *Neurocomputing*, 281:78–85, 2018.

Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. Rgb-t object tracking: Benchmark and baseline. *Pattern Recognition*, 96:106977, 2019.

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.

Hao Li, Yuhao Wang, Wenning Hao, Pingping Zhang, Dong Wang, and Huchuan Lu. Ragtrack: Language-aware rgbt tracking with retrieval-augmented generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28179–28189, 2026a.

Hao Li, Yuhao Wang, Xiantao Hu, Wenning Hao, Pingping Zhang, Dong Wang, and Huchuan Lu. Cadtrack: Learning contextual aggregation with deformable alignment for robust rgbt tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6109–6117, 2026b.

Jiacheng Li and Songhe Feng. Bridging modalities via progressive re-alignment for multimodal test-time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 22931–22939, 2026.

Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133(1):31–64, 2025.

Hyesu Lim, Byeonggeun Kim, Jaegul Choo, and Sungha Choi. Ttn: A domain-shift aware batch normalization in test-time adaptation. *arXiv preprint arXiv:2302.05155*, 2023.

Wei Lin, Muhammad Jehanzeb Mirza, Mateusz Kozinski, Horst Possegger, Hilde Kuehne, and Horst Bischof. Video test-time adaptation for action recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22952–22961. IEEE, 2023.

Andong Lu, Ziyi Zha, Jiandong Jin, Shihao Li, Chenglong Li, Jin Tang, and Bin Luo. Spatio-temporal conditional denoising transformer for modality-missing rgbt tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13584–13593, 2026.

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.

Chaithanya Kumar Mummadi, Robin Hutmacher, Kilian Rambach, Evgeny Levinkov, Thomas Brox, and Jan Hendrik Metzen. Test-time adaptation to distribution shift by confidence maximization and input transformation. *arXiv preprint arXiv:2106.14999*, 2021.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pp. 16888–16905. PMLR, 2022.

Hyoungseob Park, Anjali Gupta, and Alex Wong. Test-time adaptation for depth completion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20519–20529. IEEE, 2024.

Xiaozhen Qiao, Da Zhang, Yubin Guo, Junyu Gao, Zhiyuan Zhao, and Xuelong Li. Semantic-aware temporal adaptation for uav anti-uav tracking. *arXiv preprint arXiv:2607.26511*, 2026.

Arindam Saha, Bibhas Chandra Dhara, Saiyed Umer, Kulakov Yurii, Jazem Mutared Alanazi, and Ahmad Ali AlZubi. Efficient obstacle detection and tracking using rgb-d sensor data in dynamic environments for robotic applications. *Sensors*, 22(17):6537, 2022.

Zekai Shao, Yufan Hu, Bin Fan, and Hongmin Liu. Pura: Parameter update-recovery test-time adaption for rgb-t tracking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22089–22098, 2025.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022.

Junbin Su, Ziteng Xue, Shihui Zhang, Kun Chen, Weiming Hu, and Zhipeng Zhang. Seatrack: Simple, efficient, and adaptive multimodal tracker. *arXiv preprint arXiv:2604.12502*, 2026.

Yuedong Tan, Jiawei Shao, Eduard Zamfir, Ruanjun Li, Zhaochong An, Chao Ma, Danda Paudel, Luc Van Gool, Radu Timofte, and Zongwei Wu. What you have is what you track: Adaptive and robust multimodal tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3455–3465, 2025a.

Yuedong Tan, Zongwei Wu, Yuqian Fu, Zhuyun Zhou, Guolei Sun, Eduard Zamfir, Chao Ma, Danda Paudel, Luc Van Gool, and Radu Timofte. Xtrack: Multimodal training boosts rgb-x video object trackers. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5734–5744. IEEE, 2025b.

Zhangyong Tang, Tianyang Xu, Xiaojun Wu, Xue-Feng Zhu, and Josef Kittler. Generative-based fusion mechanism for multi-modal tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5189–5197, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7191–7201. IEEE, 2022.

Shilei Wang, Pujian Lai, Dong Gao, Jifeng Ning, and Gong Cheng. Exploring modality-aware fusion and decoupled temporal propagation for multi-modal object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10065–10073, 2026.

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(3):1997–2010, 2023.

Zixin Wang, Yadan Luo, Liang Zheng, Zhuoxiao Chen, Sen Wang, and Zi Huang. In search of lost online test-time adaptation: A survey. *International Journal of Computer Vision*, 133(3): 1106–1139, 2025.

Zongwei Wu, Jilai Zheng, Xiangxuan Ren, Florin-Alexandru Vasluianu, Chao Ma, Danda Pani Paudel, Luc Van Gool, and Radu Timofte. Single-model and any-modality for video object tracking. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19156–19166. IEEE, 2024.

Song Yan, Jinyu Yang, Jani Käpylä, Feng Zheng, Aleš Leonardis, and Joni-Kristian Kämäräinen. Depthtrack: Unveiling the power of rgbd tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10725–10733, 2021.

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.

Ram Zaveri, Shivang Patel, Yu Gu, and Gianfranco Doretto. Improving accuracy and generalization for efficient visual tracking. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 9450–9460, 2025.

Chunhui Zhang, Li Liu, Hao Wen, Xi Zhou, and Yanfeng Wang. Awesome multi-modal object tracking. *arXiv preprint arXiv:2405.14200*, 2024a.

Haiping Zhang, Di Yuan, Xiu Shu, Zhihui Li, Qiao Liu, Xiaojun Chang, Zhenyu He, and Guangming Shi. A comprehensive review of rgbt tracking. *IEEE Transactions on Instrumentation and Measurement*, 73:1–23, 2024b.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pp. 493–510. Springer, 2022.

Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9516–9526, 2023.

A Appendix for Additional Analysis of S^{2}TTA

A.1 Multi-Level Statistics and Fisher-Based Weighting

In our default setting, we construct the template prior using multi-level feature statistics and assign each level a decision-aware weight, since representations at different levels may contribute differently to target prediction. Following Hu et al. (2026), we adopt Fisher-based weighting to emphasize more decision-sensitive representations when aggregating multi-level statistics. We further investigate the effects of multi-level statistics and different weighting strategies in template prior construction. As reported in Table 4, we compare the baseline with four variants: single-level statistics extracted from the final encoder layer, multi-level statistics with uniform averaging, multi-level statistics with progressively increasing weights, and multi-level statistics with Fisher-based weighting. The single-level and uniformly weighted multi-level variants exhibit performance fluctuations across different benchmarks, indicating that simply incorporating more feature levels does not necessarily yield consistent improvements. In contrast, progressive weighting produces more stable gains, suggesting that multi-level statistics are beneficial when the contributions of different feature levels are properly differentiated, with deeper representations generally providing more informative cues for template-prior construction. Fisher-based weighting further achieves the best overall performance across all benchmarks, demonstrating that parameter-sensitivity-aware weighting can more effectively identify informative feature levels and construct a more reliable template prior. These results indicate that Fisher weighting is not a prerequisite for the effectiveness of the proposed framework, but serves as a key mechanism for fully exploiting multi-level representations and strengthening the resulting adaptation objective.

Table 4: Further analysis of multi-level feature statistics and different weighting strategies for template prior construction. We compare single-level statistics with multi-level statistics using uniform, progressively increasing, and Fisher-based weighting.

Method	RGB-T				RGB-D						RGB-E	
	LasHeR		RGBT234		DepthTrack			VOT-RGBD22			VisEvent	
	PR	AUC	MPR	MSR	Pr	Re	F-score	EAO	Acc	Rob	SR	PR
Baseline	74.8	60.0	93.5	70.2	65.1	66.3	65.7	77.3	83.1	92.8	65.0	82.3
Single-Level	74.6	59.7	93.7	71.9	62.8	63.8	63.3	75.5	81.3	90.9	65.1	82.2
Multi-Level + Uniform Weighting	74.2	59.6	94.0	72.0	64.0	64.3	64.1	76.8	82.8	92.5	65.0	82.0
Multi-Level + Progressive Weighting	75.2	60.4	94.2	72.2	65.5	66.4	65.8	77.5	83.3	93.0	65.1	82.3
Multi-Level + Fisher Weighting (Ours)	**76.0**	**60.8**	**94.8**	**72.8**	**66.2**	**66.8**	**66.5**	**77.9**	**83.5**	**93.2**	**65.5**	**82.6**

A.2 Efficiency Analysis

We further evaluate the runtime efficiency of S^{2}TTA by comparing its inference speed with the baseline and several representative TTA methods, including Tent Wang et al. (2020), EATA Niu et al. (2022), PURA Shao et al. (2025), and IDEA Hu et al. (2026). As reported in Table 5, all TTA methods introduce additional computational overhead compared with the baseline. Moreover, S^{2}TTA achieves substantially higher FPS than EATA and IDEA, mainly because it avoids repeated backward or forward passes during online adaptation. These results indicate that S^{2}TTA maintains a favorable balance between adaptation effectiveness and computational efficiency, while introducing only moderate additional computational overhead.

Table 5: Runtime efficiency comparison with representative TTA methods. FPS is reported for the baseline, Tent, EATA, PURA, IDEA, and S^{2}TTA.

Method	RGB-T						RGB-D								RGB-E			Avg. FPS
	LasHeR			RGBT234			DepthTrack				VOT-RGBD22				VisEvent			
	PR	AUC	FPS	MPR	MSR	FPS	Pr	Re	F-score	FPS	EAO	Acc	Rob	FPS	SR	PR	FPS	
Baseline	74.8	60.0	25.0	93.5	70.2	24.2	65.1	66.3	65.7	27.9	77.3	83.1	92.8	25.7	65.0	82.3	28.7	26.3
Tent	74.5	59.5	24.0	94.1	71.0	23.1	64.8	65.9	65.5	22.0	76.9	82.6	92.4	22.1	64.7	81.6	23.8	23.0
EATA	74.6	59.8	6.8	94.2	71.0	7.0	64.7	65.7	65.2	6.3	77.0	82.8	92.4	7.0	64.8	81.8	7.3	6.9
PURA	75.0	60.2	24.3	93.8	70.6	23.1	65.3	66.2	65.9	23.7	77.1	82.9	92.6	20.0	65.0	82.2	22.3	22.7
IDEA	74.8	60.0	2.3	93.3	70.2	3.0	65.4	63.7	62.9	2.3	77.4	83.3	92.9	2.3	65.1	82.3	2.2	2.4
S^{2}TTA (Ours)	**76.0**	**60.8**	11.9	**94.8**	**72.8**	12.2	**66.2**	**66.8**	**66.5**	12.0	**77.9**	**83.5**	**93.2**	11.7	**65.5**	**82.6**	10.7	11.7

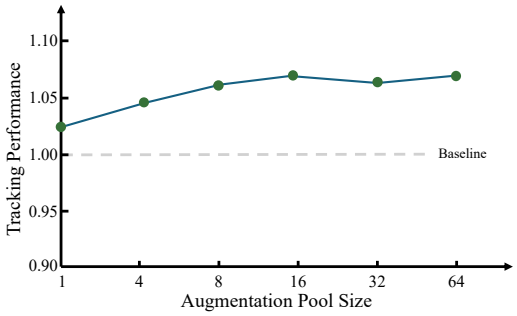


Figure 4: Analysis of the template augmentation pool size. Performance is reported across different pool sizes, while the dashed line denotes the baseline tracker without S^{2}TTA.

A.3 Analysis of Template Augmentation Pool Size

We construct the template augmentation pool using geometric transformations and color perturbations, with a default pool size of 16. We further analyze its effect on tracking performance. As shown in Fig. 4, performance generally improves as the pool size increases up to 16, where the best performance is achieved, while further increasing the pool size leads to moderate fluctuations without consistent gains. Notably, even without additional augmentation, S^{2}TTA still outperforms the baseline tracker, indicating that its effectiveness primarily stems from the target-specific template-guided adaptation rather than data augmentation. Introducing augmented template views further enriches target variations and improves the robustness of the constructed prior. However, excessively large pools provide diminishing benefits and may introduce redundant or overly perturbed samples, resulting in mild performance fluctuations.

B Limitations and Future Work

B.1 Limitations

By exploiting the inherent target template as a target-specific adaptation reference, S^{2}TTA offers a new perspective for test-time adaptation in multimodal object tracking. Despite its consistent robustness across different multimodal tracking settings, several limitations remain. First, the current uncertainty-aware prior evolution primarily relies on response entropy to assess prediction reliability. Although entropy provides an efficient uncertainty measure, highly confident but incorrect predictions may still pass the uncertainty gate under challenging target ambiguities or severe distribution shifts. Second, the current selective online adaptation uses fixed thresholds and a fixed momentum coefficient to ensure stable online decisions. This design is effective in practice, but leaves room for more sequence-adaptive control. Moreover, online prompt optimization inevitably introduces additional computational overhead compared with the frozen baseline, despite being substantially more efficient than iterative TTA approaches.

B.2 Future Work

Future work could first explore more reliable uncertainty estimation for prior evolution, either by jointly modeling response uncertainty, temporal consistency, cross-modal agreement, and historical prediction behavior, or by introducing new mechanisms for assessing prediction reliability. Another important direction is to develop more adaptive and flexible shift-aware mechanisms that can better characterize varying distribution shifts across modalities and sequences, enabling more precise selection of adaptation spaces and update intensity. From an efficiency perspective, future work could further improve the computational efficiency of online adaptation without compromising its flexibility and adaptation capability. More broadly, a promising direction is to develop a more unified and continuous adaptation mechanism that jointly models distribution shift, prediction uncertainty, and adaptation intensity for more reliable long-term test-time adaptation.