Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

CCLT: Correction-Conditioned Low-Rank Exchange for Parameter-Efficient RGB-X Tracking

Xingyu Tan
Beijing University of Posts and Telecommunications
Chun Liu
Beijing University of Posts and Telecommunications
Qing Song
Beijing University of Posts and Telecommunications
Mengjie Hu
18011248857@163.com

Beijing University of Posts and Telecommunications

Research Article

Keywords:

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11441577/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

CCLT: Correction-Conditioned Low-Rank Exchange for Parameter-Efficient RGB-X Tracking

Xingyu Tan[1], Chun Liu[1], Qing Song[1], Mengjie Hu[1,*]

[1]School of Intelligent Engineering and Automation, Beijing University of Posts and Telecommunications, Beijing 100876, China

*Corresponding author: Mengjie Hu, mengjie.hu@bupt.edu.cn

ORCID:
Xingyu Tan: 0009-0005-3363-2121
Chun Liu: 0000-0002-2834-9461
Qing Song: 0000-0003-4616-2200
Mengjie Hu: 0000-0001-7712-3322

Abstract

Parameter-efficient fine-tuning (PEFT) offers a practical way to adapt pretrained RGB trackers to RGB-X tracking, where paired multimodal training data are limited. Despite their architectural diversity, existing PEFT approaches typically fuse the RGB and X branches through two parallel directional corrections. The $(X \to \mathrm{RGB})$ correction updates the pre-fusion RGB state using the pre-fusion X state, while the $(\mathrm{RGB} \to X)$ correction updates the pre-fusion X state using the pre-fusion RGB state. This parallel formulation prevents the $(\mathrm{RGB} \to X)$ correction from using the corrected RGB state produced by the $(X \to \mathrm{RGB})$ correction. However, the source-state dependency between the two directional corrections remains underexplored. To model this source-state dependency under a compact trainable budget, we introduce Correction-Conditioned Low-rank Tracking (CCLT), a parameter-efficient RGB-X tracker that combines sparsely inserted bidirectional Lift exchange (BiLift) modules with lightweight Intra-Stream LoRA (IntraLoRA) adapters. BiLift applies an $(X \to \mathrm{RGB} \to X)$ corrective relay at one selected backbone depth and an $(\mathrm{RGB} \to X \to \mathrm{RGB})$ relay at another, allowing the second directional correction at each depth to read a freshly corrected source state. IntraLoRA reuses low-rank key/value factors across the RGB and X streams to adapt within-stream representations while keeping the shared backbone frozen, thereby balancing cross-modal exchange with parameter-efficient representation adaptation. Together, these innovations enable CCLT to achieve a more favorable balance between tracking performance and computational efficiency than existing state-of-the-art methods across RGB-T, RGB-D and RGB-E tracking tasks.

Keywords: RGB-X tracking, Parameter-efficient fine-tuning, Cross-modal interaction, Low-rank adaptation, Visual object tracking

1 Introduction

Visual object tracking localizes a target throughout a video from its initial appearance. RGB trackers remain vulnerable when illumination changes or rapid motion make appearance ambiguous. RGB-X tracking improves robustness by pairing RGB frames with auxiliary thermal, depth or event observations. Yet paired RGB-X annotations are scarce relative to RGB pretraining data [1–3]. Full fine-tuning (FFT) is computationally costly and risks catastrophic forgetting, whereas parameter-efficient fine-tuning (PEFT) freezes pretrained tracking knowledge and learns only a small set of modality-adaptation parameters [1, 2, 4].

Existing PEFT RGB-X trackers follow two interaction paradigms. Auxiliary-conditioned prompting (Fig. 1(a)) injects X-derived prompts into an RGB-centric foundation [1, 4]. However, frozen RGB-pretrained weights are poorly matched to mixed-modality inputs, inducing distribution shifts and erroneous attention shifts [5]. This reinforces SDSTrack's conclusion [2] that one-stream designs are less robust than two-stream alternatives. Reciprocal two-stream interaction (Fig. 1(b)) preserves separate RGB and X states and exchanges information bidirectionally through adapters, symmetric adaptation, low-rank prompting or attention guidance. However, these methods overlook state progression across the two directions. Existing designs place two one-way operators side by side and compute both from the same uncorrected states, so neither can exploit the other's correction. The reverse direction therefore duplicates a one-hop update rather than producing state-progressive interaction [2, 5–7].

Existing methods specify where information flows but fail to model how the two corrections compose. Thus, an urgent core question is how to introduce state dependency into bidirectional cross-modal exchange under a fixed parameter budget, enabling the later correction to build on the state produced by the earlier one.

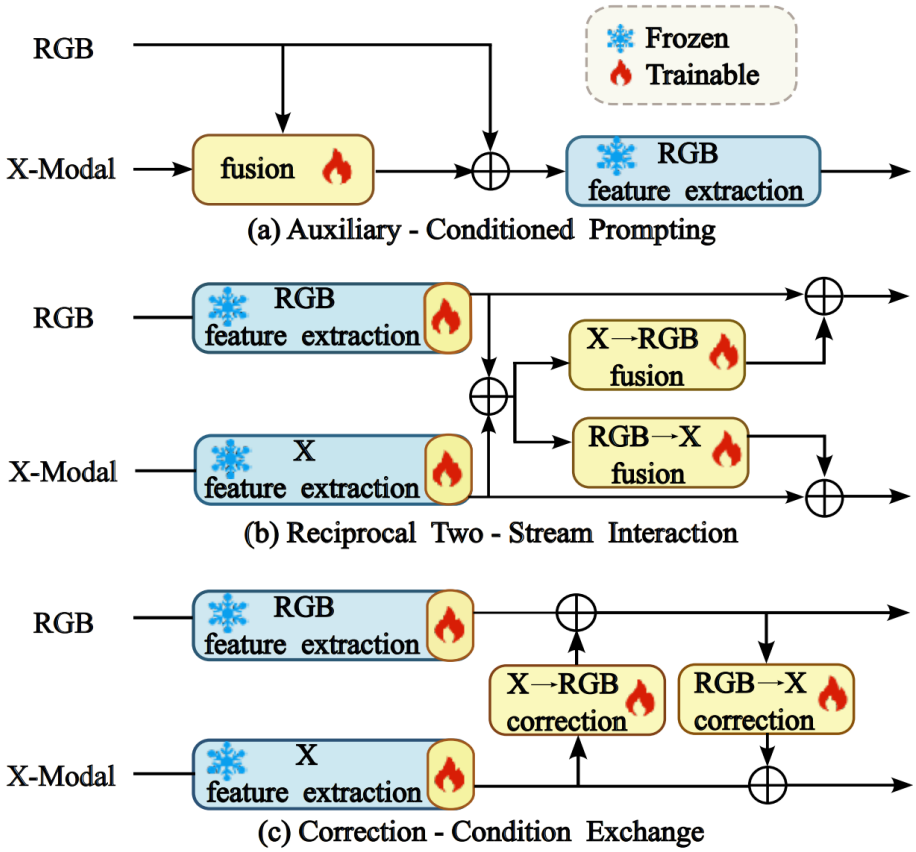


Figure 1 Interaction paradigms in parameter-efficient RGB-X tracking. (a) Auxiliary-conditioned prompting [1] injects X-modal cues into an RGB-centric stream. (b) Reciprocal two-stream interaction [5] maintains separate RGB and X paths and enables bidirectional correction. (c) CCLT conditions the second BiLift correction on the state updated by the first

To address this question, we propose CCLT, a parameter-efficient RGB-X tracker that combines sparsely inserted correction modules with lightweight adapters on a shared frozen two-stream backbone, as illustrated in Fig. 1(c). Its two components serve complementary roles: bidirectional Lift exchange (BiLift) models correction-conditioned cross-modal interaction, whereas IntraLoRA provides modality-shared within-stream adaptation. Specifically, at one selected depth, BiLift forms an $(X \to \mathrm{RGB} \to X)$ corrective relay: the $X \to \mathrm{RGB}$ branch first applies a low-rank residual to the RGB state, and the $\mathrm{RGB} \to X$ branch then reads the corrected RGB state to update X. At the other depth, BiLift reverses the order to $(\mathrm{RGB} \to X \to \mathrm{RGB})$, allowing each modality to initiate one relay while making the second directional correction depend explicitly on the first. IntraLoRA reuses key-specific and value-specific low-rank factors across modalities while adapting the two streams independently. To isolate corrected-state dependency, we construct Parallel-LR, which matches BiLift in directional operator design, insertion sites and trainable parameters but computes both corrections from the incoming states.

With all inherited parameters frozen, CCLT updates only 196.6K parameters (0.212% of the tracker). Under a parameter-matched three-seed RGB-T protocol, it improves mean LasHeR SR over Parallel-LR by $+1.980$ points (95% interval: $[0.360, 2.690]$), supporting corrected-state dependency. Across five RGB-T/RGB-D/RGB-E benchmarks, it maintains comparable performance while delivering about 68.2% higher FPS than comparable methods.

Overall, our contributions are:

- **CCLT.** We propose CCLT, a simple and efficient tracker that combines lightweight low-rank adapters for parameter-efficient adaptation with sparsely inserted correction modules for cross-modal interaction. It achieves a state-of-the-art performance–efficiency trade-off and positions sparse correction as a promising direction for lightweight multimodal fusion.
- **BiLift.** We target the underexplored state dependency in PEFT RGB-X interaction, by proposing correction-conditioned low-rank exchange with BiLift. BiLift organizes the two cross-modal low-rank corrections into a sequential corrective relay. Extensive experiments show its strong potential as a reliable baseline, offering a new perspective on inter-modal interaction.
- **IntraLoRA.** To address paired-data scarcity, tuning cost and adaptation–interaction coordination, we propose IntraLoRA. It reuses parameters for within-stream adaptation, balancing multimodal adaptation and

cross-modal exchange. It achieves 68.2% higher FPS than comparable methods while maintaining comparable performance.

2 Related Work

2.1 Parameter-Efficient RGB-X Tracking

PEFT adapts a pretrained model through a small trainable interface while keeping the foundation weights frozen. Common interfaces include visual prompts, bottleneck adapters and low-rank updates [8–10]. This setting is particularly relevant to RGB-X tracking, where paired multimodal annotations are limited relative to RGB pretraining data [1, 2], making parameter-efficient adaptation especially desirable.

Existing trackers realize PEFT through different combinations of within-stream adaptation and cross-stream communication. LoRAT studies LoRA for RGB tracking, while ViPT uses auxiliary-conditioned prompts; BAT, SDSTrack and LRPD introduce bidirectional adapters, masked cross-modal attention and symmetric LoRA prompting, respectively [1, 2, 6, 7, 11]. XTrack, GLoRA and SEATrack further combine low-rank adaptation with shared or guided multimodal exchange, whereas SDMoEA and PATrack organize expert or progressive modality-specific interfaces [5, 12–15]. CCLT differs in the composition of its two directional corrections rather than in adding another adapter family.

2.2 Cross-Modal Interaction

Cross-modal interaction spans input-level, feature-level and attention-level designs. ProTrack maps aligned modalities to one prompted image and ViPT conditions an RGB stream with auxiliary-generated prompts [1, 16]. Two-stream trackers instead exchange features through reciprocal adapters or attention, as in BAT, SDSTrack, LRPD and UBPT, while TBSI, IIMF, CAFormer and SEATrack introduce template–search interaction, shared representations, cross-modulated attention or guided expert fusion [2, 5–7, 17–20].

The closest designs establish two-way communication, but the labels *symmetric* and *bidirectional* do not specify which state a second update reads. BAT transfers corresponding source-stream features, SEATrack's AMG jointly guides two LoRA-enhanced attention maps and SDSTrack and LRPD couple the streams through masked attention or symmetric prompting. None of these studies isolates whether the reverse update reads the incoming state or the state corrected by the first update under a parameter-matched control. Ordered additive coupling is established in reversible networks and paired decomposition streams [21, 22]; CCLT does not claim that algebra as new. Instead, it instantiates corrected-state dependency as sparse low-rank RGB-X exchange and evaluates that dependency against a matched parallel alternative.

3 Method

We present CCLT, an efficient RGB-X tracker built on a frozen ViT-Base tracking foundation [23]. It combines IntraLoRA (Intra-Stream LoRA), a modality-shared configuration of key/value LoRA [10], with BiLift for correction-conditioned low-rank exchange. Figure 2 links the four scales used below: end-to-end dataflow, encoder, block and module internals. Accordingly, the following subsections describe the overall architecture, IntraLoRA, BiLift and optimization in that order.

3.1 Overall Architecture

As shown on the left of Figure 2, CCLT keeps synchronized RGB and X inputs in separate streams while sharing the inherited tokenization, backbone and prediction operators. Let (z^R, z^X) and (x_t^R, x_t^X) denote the template and search pairs at time t, where X is thermal, depth or event data. Following the one-stream formulation [23], both modalities share patch embedding $\mathscr{E}$ and positional embeddings p_z, p_x. For $m \in \{R, X\}$,

$$H_m^{(0)} = \left[\mathscr{E}(z^m) + p_z\,;\, \mathscr{E}(x_t^m) + p_x\right] \tag{1}$$

where $[\,;\,]$ denotes concatenation, yielding distinct states $H_R^{(0)}$ and $H_X^{(0)}$.

The $L = 12$ shared Transformer blocks map $H_m^{(\ell-1)}$ to $H_m^{(\ell)}$ through the same frozen sequence. IntraLoRA adapts each stream with modality-shared factors, while BiLift exchanges corrections at blocks 6 and 10. The shared final LayerNorm is applied per stream; the normalized search tokens are then summed for the frozen center head. Only IntraLoRA factors and BiLift projections are optimized.

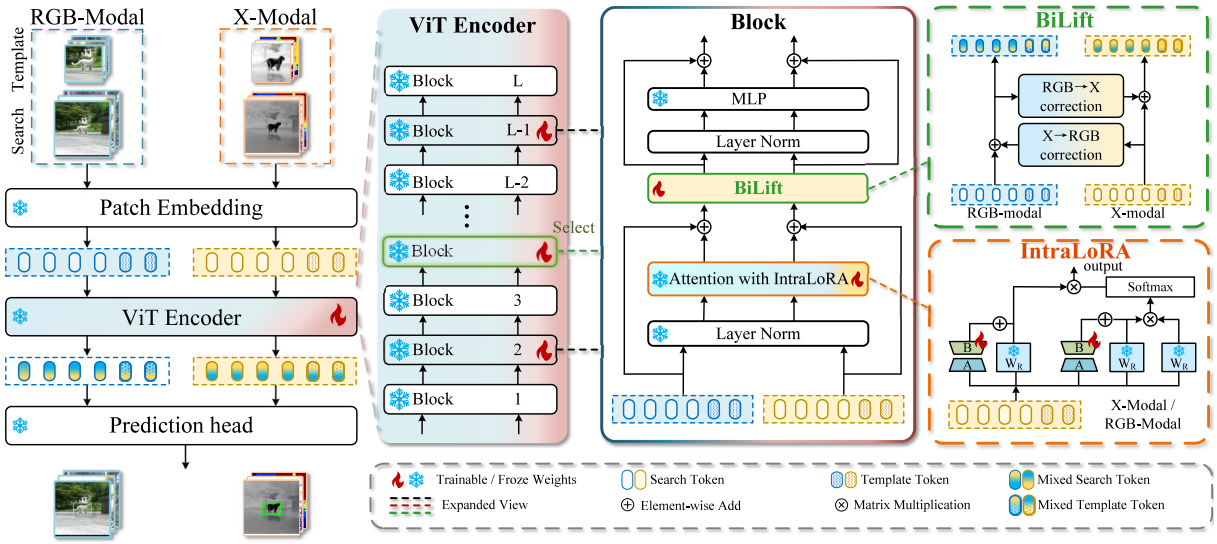


Figure 2 Overview of CCLT from left to right. RGB-X template–search pairs are processed by a shared tracking backbone, whose 12-block ViT encoder is expanded next. IntraLoRA is inserted into every even-numbered block; the green highlights mark blocks 6 and 10, where BiLift is placed. The block-level view shows the module placements. The rightmost insets detail BiLift's correction-conditioned exchange and IntraLoRA's low-rank key/value adaptation

3.2 IntraLoRA: Within-Stream Adaptation

The lower-right inset of Figure 2 locates IntraLoRA within self-attention. Following SEATrack [5], IntraLoRA applies LoRA [10] to the key and value projections of selected self-attention blocks:

$$\Delta W_p^{(\ell)} = B_p^{(\ell)} A_p^{(\ell)} \tag{2}$$

For projection $p \in \{k, v\}$ at block ℓ, $A_p^{(\ell)}$ has shape $r \times d$ and projects the d-channel state to rank r, whereas $B_p^{(\ell)}$ has shape $d \times r$ and maps it back; we use $d = 768$ and $r = 8$. Given $H_m^{(\ell)}$, the projection sums the frozen branch $H_m^{(\ell)}(W_p^{(\ell)})^\top$ and the low-rank branch $\mathrm{Drop}_{0.1}\left(H_m^{(\ell)}(\Delta W_p^{(\ell)})^\top\right)$, with dropout disabled at inference. Within these attention projections, only $A_p^{(\ell)}$ and $B_p^{(\ell)}$ are optimized; the query projection and pretrained K/V weights remain frozen. We add BA directly, without an additional α/r multiplier.

We apply IntraLoRA at every other Transformer block, starting from the second; the adapted blocks are $\ell \in \{2,4,6,8,10,12\}$. At each adapted block, IntraLoRA maintains separate factor pairs for the key and value projections and reuses each pair across the RGB and X streamwise evaluations; different blocks retain distinct factors. Because the shared attention operator is evaluated separately on the two streams, each IntraLoRA update reads only its own current state and therefore introduces no cross-stream state dependency. IntraLoRA thus provides within-stream adaptation, while BiLift supplies the corrected-state dependency described next.

3.3 BiLift: Corrected-State Low-Rank Exchange

Figure 2 locates BiLift in the backbone, while Fig. 3 details its two update orders and directional low-rank correction. Bidirectional and symmetric cross-modal adaptation has been explored through adapters and LoRA prompting in parameter-efficient multimodal tracking [2, 6, 7]. These designs motivate adapting both modality streams, but bidirectionality or symmetry alone does not specify whether a directional correction reads an incoming state or a state corrected earlier within the same interaction site. To make this source-state choice explicit under a fixed low-rank operator form, we define a parallel reference.

Let R_ℓ and X_ℓ denote the post-attention RGB and X states entering an interaction site at block ℓ, respectively, and let $\mathscr{F}_{a\to b}^{(\ell)}$ denote a directional residual transform from source stream a to target stream b. The parallel reference computes

$$\begin{aligned} R_\ell^{\parallel} &= R_\ell + \mathscr{F}_{X\to R}^{(\ell)}(X_\ell) \\ X_\ell^{\parallel} &= X_\ell + \mathscr{F}_{R\to X}^{(\ell)}(R_\ell) \end{aligned} \tag{3}$$

We term this formulation Parallel-LR. Because both directional transforms read the site-input states, neither correction can depend on the correction produced by the other within the same site. Parallel-LR therefore represents

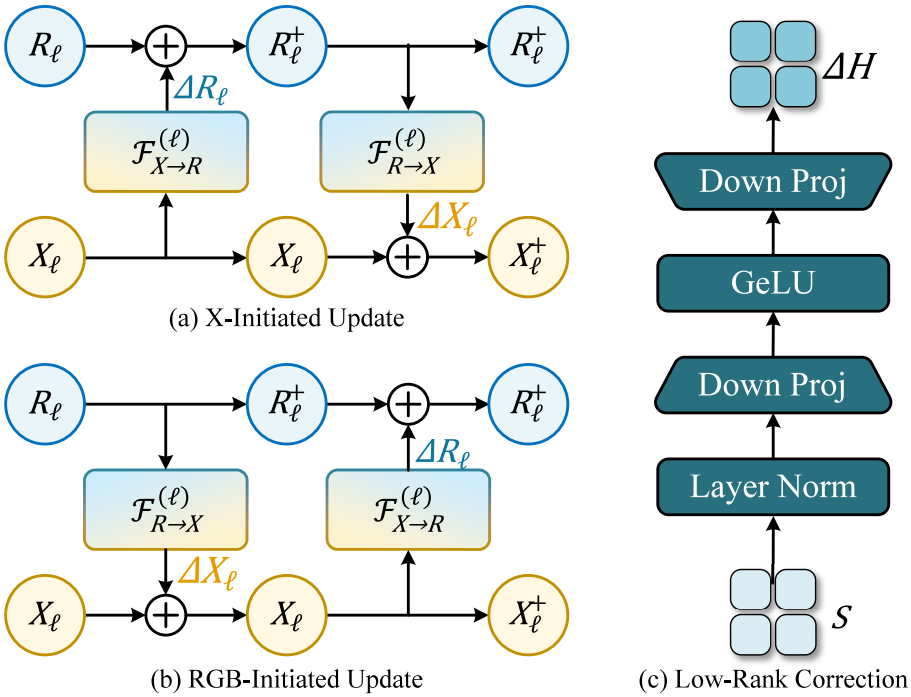


Figure 3 BiLift update order and correction-conditioned directional low-rank exchange mechanism

bidirectional exchange without within-site correction composition.

Corrected-state dependency. BiLift introduces corrected-state dependency while retaining the directional low-rank operator form of Parallel-LR. For the X-initiated orientation in Fig. 3(a), BiLift computes

$$\begin{aligned} R_\ell^+ &= R_\ell + \mathscr{F}^{(\ell)}_{X\to R}(X_\ell) \\ X_\ell^+ &= X_\ell + \mathscr{F}^{(\ell)}_{R\to X}(R_\ell^+) \\ &= X_\ell + \mathscr{F}^{(\ell)}_{R\to X}\Big(R_\ell + \mathscr{F}^{(\ell)}_{X\to R}(X_\ell)\Big) \end{aligned} \tag{4}$$

The sequential update in Eq. (4) is algebraically equivalent to the additive coupling of reversible residual networks [21]. Here, the coupled states are synchronized RGB and X representations and the directional functions are low-rank cross-modal corrections. Writing $\Delta R_\ell = \mathscr{F}^{(\ell)}_{X\to R}(X_\ell)$ and $\Delta X_\ell = \mathscr{F}^{(\ell)}_{R\to X}(R_\ell^+)$, Fig. 3(a) exposes the order $\Delta R_\ell \to R_\ell^+ \to \Delta X_\ell \to X_\ell^+$: the second branch reads R_ℓ^+, which already contains ΔR_ℓ. This corrected-state dependency, rather than bidirectionality alone, is the defining mechanism of BiLift. The Section 3.4 states and proves a site-level proposition comparing BiLift with Parallel-LR: it identifies the corrected-state composition term, gives the explicit additive-coupling inverse and unit Jacobian-determinant result, bounds the composed path by the low-rank bottleneck and analyzes the complementary orientation.

Directional low-rank correction. As expanded in Fig. 3(c), each $\mathscr{F}^{(\ell)}_{a\to b}$ maps a source state S of shape $n_\ell \times d$ to an additive residual of the same shape:

$$\mathscr{F}^{(\ell)}_{a\to b}(S) = \mathrm{GELU}\Big(\mathrm{LN}_0(S)\,(D^{(\ell)}_{a\to b})^\top\Big)\,(U^{(\ell)}_{a\to b})^\top \tag{5}$$

Here, LN_0 is an affine-free LayerNorm, while $D^{(\ell)}_{a\to b}$ and $U^{(\ell)}_{a\to b}$ are bias-free down- and up-projections of shapes $r\times d$ and $d\times r$, respectively. We use $d = 768$ and $r = 8$, with no dropout in the evaluated configuration. The two directions at each site use separate transforms and projection parameters are also distinct across sites. We use the term *correction* only for this learned additive low-rank residual; it does not denote an explicitly supervised error signal.

At each interaction site, BiLift operates on the post-attention residual states and returns the corrected pair before candidate elimination and the MLP sublayer. As shown in Fig. 3(b), exchanging the roles of R and X in Eq. (4) gives the complementary RGB-initiated orientation, $X_\ell^+ = X_\ell + \mathscr{F}^{(\ell)}_{R\to X}(R_\ell)$ followed by $R_\ell^+ = R_\ell + \mathscr{F}^{(\ell)}_{X\to R}(X_\ell^+)$. Reversing the initiating stream changes which modality receives the within-site two-hop dependency while preserving the same directional low-rank operators. CCLT alternates the orientations in Fig. 3(a,b) across its two BiLift sites, so each modality initiates one corrective relay.

3.4 Site-Level Structural Analysis of BiLift

At one interaction site, we ask what changes when the second directional correction reads the state produced by the first rather than the same incoming state used by a parallel update. We isolate this source-state choice by comparing BiLift with a parameter-matched Parallel-LR topology that uses the same directional correction functions. Proposition 1 answers this question. It proves that BiLift introduces the within-site two-hop composition term B^+A, whereas Parallel-LR has no corresponding within-site term. The proof gives an explicit inverse and unit Jacobian determinant for the ordered BiLift map and bounds the composed path by the directional low-rank bottlenecks. These results establish BiLift's structural advantage: without adding directional operators or trainable parameters relative to Parallel-LR, its second correction responds to a freshly corrected source and thereby realizes an additional within-site corrected-state dependency path.

This advantage arises from a simple but consequential change in computation order rather than from a larger module. BiLift reorganizes the same two low-rank directional corrections into a state-progressive relay. The additive-coupling algebra of reversible residual networks (RevNets) [21] provides a compact tool for characterizing this relay. Applied to BiLift, it connects the implementation-level reordering to the composition, inverse, determinant, and rank properties established below. The controlled experiments in Sect. 4 then evaluate how this structural advantage translates into tracking performance.

3.4.1 Definitions and Assumptions

Vectorize the two incoming token states as $R \in \mathbb{R}^m$ and $X \in \mathbb{R}^n$. Let $f : \mathbb{R}^n \to \mathbb{R}^m$ denote the X-to-RGB correction and let $g : \mathbb{R}^m \to \mathbb{R}^n$ denote the RGB-to-X correction. We make the following assumptions.

A1 Fixed deterministic site. The parameters of f and g are fixed during one forward map, their outputs have the residual-compatible dimensions above, and repeated evaluation at the same input returns the same output. In the evaluated checkpoints, dropout in these branches is zero.

A2 Local differentiability. For the Jacobian statements, f and g are continuously differentiable in neighborhoods of the states at which their derivatives are evaluated. This assumption is not needed for the explicit inverse.

A3 Topology-isolated comparison. To compare state dependencies, the same arbitrary functions f and g are inserted into both maps below; only the state read by the second branch changes. This is an equation-level control and does not assert that separately trained BiLift and Parallel-LR checkpoints learn identical weights.

A4 Tokenwise low-rank branch. Only for the rank bound, each directional correction acts independently on N tokens of width d as

$$\begin{aligned} F(s) &= U\,\sigma\big(D\,\mathrm{LN}(s)\big) \\ D \in \mathbb{R}^{r\times d}, &\qquad U \in \mathbb{R}^{d\times r} \end{aligned} \tag{6}$$

where σ is GELU. This is the column-vector form of the implementation and main-paper equation.
For the X-initiated orientation, define the corrected-state BiLift map

$$\begin{aligned} T_B(R,X) &= (R^+, X^+) \\ R^+ &= R + f(X) \\ X^+ &= X + g(R^+) \end{aligned} \tag{7}$$

and the parallel control

$$\begin{aligned} T_P(R,X) &= (R^{\parallel}, X^{\parallel}) \\ R^{\parallel} &= R + f(X) \\ X^{\parallel} &= X + g(R) \end{aligned} \tag{8}$$

3.4.2 Formal Comparison

Proposition 1. (Corrected-state composition, reversibility, and rank). *Under A1–A3, the maps in Eqs.* (7) *and* (8) *have the following properties.*

1. If A2 holds and

$$A = Df(X), \qquad B^+ = Dg(R^+), \qquad B^0 = Dg(R) \tag{9}$$

then, with inputs and outputs ordered as (R, X)*,*

$$J_B = \begin{bmatrix} I_m & A \\ B^+ & I_n + B^+ A \end{bmatrix} \tag{10}$$

$$J_P = \begin{bmatrix} I_m & A \\ B^0 & I_n \end{bmatrix} \tag{11}$$

Thus B^+A *is the differential contribution of the within-site path* $X \to R^+ \to X^+$*. The corresponding diagonal block of* J_P *is* I_n *and contains no such two-hop term.*

2. T_B *is bijective even if neither* f *nor* g *is invertible. Its inverse is*

$$\begin{aligned} X &= X^+ - g(R^+) \\ R &= R^+ - f(X) \end{aligned} \tag{12}$$

If A2 also holds, $\det J_B = 1$*.*

3. *In contrast, the parallel topology is not unconditionally invertible:*

$$\det J_P = \det\left(I_n - B^0 A\right) \tag{13}$$

which can be zero.

4. *Under A4, the Jacobian of either tokenwise directional branch has rank at most* Nr *after vectorizing all* N *tokens. Hence*

$$\operatorname{rank}(B^+A) \le \min\{\operatorname{rank}(B^+), \operatorname{rank}(A)\} \le Nr \tag{14}$$

The corrected-state path is therefore explicitly bottlenecked by the directional low-rank branches.

Proof. *Differentiate the first line of Eq.* (7)*. Its two partial derivatives are* $\partial R^+/\partial R = I_m$ *and* $\partial R^+/\partial X = A$*. The second line gives*

$$\frac{\partial X^+}{\partial R} = B^+ \frac{\partial R^+}{\partial R} = B^+ \tag{15}$$

$$\frac{\partial X^+}{\partial X} = I_n + B^+ \frac{\partial R^+}{\partial X} = I_n + B^+ A \tag{16}$$

Stacking these four blocks proves Eq. (10)*. For Eq.* (11)*,* g *reads* R *rather than* R^+*. Therefore* $\partial X^{\|}/\partial R = B^0$ *and* $\partial X^{\|}/\partial X = I_n$*, proving Eq.* (11)*. This also locates the chain-rule composition* B^+A *exclusively in the corrected-state map.*

For any output pair (R^+, X^+)*, the second update of Eq.* (7) *can first be undone because its source* R^+ *is available:* $X = X^+ - g(R^+)$*. Substituting the recovered* X *into the first update then gives* $R = R^+ - f(X)$*. These operations recover a unique input pair, and direct substitution reproduces the output pair, proving bijectivity and Eq.* (12)*. No inverse of* f *or* g *is used.*

The BiLift Jacobian factors as

$$J_B = \begin{bmatrix} I_m & 0 \\ B^+ & I_n \end{bmatrix} \begin{bmatrix} I_m & A \\ 0 & I_n \end{bmatrix} \tag{17}$$

Both factors are block-unit-triangular, so each has determinant one and $\det J_B = 1$*. These inverse and determinant conclusions are the standard additive-coupling properties used by RevNets [21].*

For the parallel Jacobian, taking the Schur complement of its identity upper-left block yields $\det J_P = \det(I_n - B^0 A)$*. Singularity is possible: in the scalar case* $m = n = 1$*, choosing* $f(x) = x$ *and* $g(r) = r$ *gives a Jacobian with determinant zero. This example establishes that the parallel form has no architecture-level invertibility guarantee analogous to the sequential coupling.*

Finally, for one token under A4, the chain rule gives

$$DF(s) = U\, J_\sigma(s)\, D J_{\mathrm{LN}}(s) \tag{18}$$

where $J_\sigma(s)$ *and* $J_{\mathrm{LN}}(s)$ *are the activation and LayerNorm Jacobians evaluated at their corresponding intermediate states. The rank of this product is at most* $\operatorname{rank}(U) \le r$*. Independent tokenwise evaluation makes the full derivative block diagonal with* N *such blocks, so its rank is at most* Nr*. Applying* $\operatorname{rank}(PQ) \le \min\{\operatorname{rank}(P), \operatorname{rank}(Q)\}$ *to* B^+A *proves Eq.* (14)*.*

3.4.3 Complementary Initiating Direction

Corollary 1. (RGB-initiated orientation). *Under the assumptions of Proposition 1, exchange the initiating stream:*

$$\begin{aligned} X^+ &= X + g(R) \\ R^+ &= R + f(X^+) \end{aligned} \tag{19}$$

Let $B = Dg(R)$ and $A^+ = Df(X^+)$. Then

$$J_{\text{RGB-init}} = \begin{bmatrix} I_m + A^+B & A^+ \\ B & I_n \end{bmatrix} = \begin{bmatrix} I_m & A^+ \\ 0 & I_n \end{bmatrix} \begin{bmatrix} I_m & 0 \\ B & I_n \end{bmatrix} \tag{20}$$

The two-hop term moves to $\partial R^+/\partial R = I_m + A^+B$, $\det J_{\text{RGB-init}} = 1$, and the inverse is

$$\begin{aligned} R &= R^+ - f(X^+) \\ X &= X^+ - g(R) \end{aligned} \tag{21}$$

Under A4, $\operatorname{rank}(A^+B) \le Nr$.

Proof. *Differentiating Eq.* (19) *gives $\partial X^+/\partial R = B$, $\partial X^+/\partial X = I_n$, $\partial R^+/\partial R = I_m + A^+B$ and $\partial R^+/\partial X = A^+$, which proves the displayed Jacobian. Its factorization into two block-unit-triangular matrices gives determinant one. Reversing the update order gives Eq.* (21)*. The rank claim follows from the same argument as in Proposition 1.*

3.4.4 Structural Advantage of Corrected-State Evaluation

Proposition 1 establishes that computation order alone changes the dependency structure of cross-modal interaction. With the same two directional low-rank functions and parameter budget as Parallel-LR, BiLift evaluates the second correction on a freshly updated state. The chain rule therefore introduces the composition term B^+A, corresponding to the within-site path $X \to R^+ \to X^+$, whereas the parallel update contains no corresponding two-hop term. BiLift consequently gains an additional composed interaction path without adding a directional operator or trainable parameter. Moreover, its derivative remains bounded by the same rank-r directional bottlenecks, preserving the efficiency of the low-rank design.

The ordered update also retains the exact inverse and unit Jacobian determinant of additive coupling, placing the richer dependency structure inside an information-preserving site map. Corollary 1 further shows that alternating the initiating direction transfers the composition term to the other stream rather than eliminating it. Across alternating sites, both modalities can therefore receive corrected-state feedback through the complementary B^+A and A^+B paths under the same low-rank and parameter constraints. Together, these results identify update order, rather than module size, as the source of BiLift's structural advantage over Parallel-LR. The parameter-matched comparison and correction diagnostics in Sects. 4.4 and 4.5 connect this structural mechanism to the observed behavior and performance of trained trackers.

3.5 Parameter and Analytical MAC Accounting

The IntraLoRA parameter count is

$$6 \times 2 \times 2 \times 768 \times 8 = 147456 \tag{22}$$

There is no stream factor in this parameter equation because the same K/V weights serve both streams. The directional Lift parameters are

$$2 \times 2 \times 2 \times 768 \times 8 = 49152 \tag{23}$$

and the audited trainable sum is

$$147456 + 49152 = 196608 \tag{24}$$

The analytical linear-projection MAC count follows the implemented graph per batch-1 template/search pair. In the training-mode unmerged graph, `MergedLinear.merge_AB` first forms a dense zero-padded delta product spanning the full QKV output and then applies that dense delta to each stream. Its two terms are the dense token–delta product and the grouped BA materialization:

$$\begin{aligned} &2 \times 6 \times 320 \times 768 \times (768 + 768 + 768) \\ &+ 2 \times 6 \times 2 \times 768 \times 8 \times 768 = 6.90801869\,\text{G} \end{aligned} \tag{25}$$

The two ordered, bidirectional Lift sites contribute

$$2 \times 2 \times 2 \times 320 \times 768 \times 8 = 0.01572864\,\mathrm{G} \tag{26}$$

so all adapter-related linear operations sum to

$$6.90801869 + 0.01572864 = 6.92374733\,\mathrm{G} \tag{27}$$

Together, these counts show that CCLT's efficiency follows from sharing IntraLoRA parameters across streams and confining BiLift exchange to two sites. Modality-shared IntraLoRA avoids duplicating K/V factors across the two streams, while the two rank-8 BiLift sites add only 49152 parameters and 0.01572864 G linear MACs for bidirectional corrected-state exchange. The complete design therefore updates 196608 parameters (0.212% of the 92.7M-parameter tracker), with BiLift accounting for only one quarter of this trainable budget. Because BiLift and parameter-matched Parallel-LR use the same directional low-rank operators, BiLift obtains the additional corrected-state dependency established in Sect. 3.4 with no increase in parameter count or analytical linear MACs over the parallel topology.

3.6 Optimization

All inherited OSTrack parameters [23], including the Transformer backbone and center head, are frozen; we optimize only $\Theta = \Theta_{\text{IntraLoRA}} \cup \Theta_{\text{BiLift}}$. BiLift up-projections are zero-initialized, so the new exchange branches initially contribute no residuals. This leaves 196.6K trainable parameters (0.212% of the 92.7M-parameter tracker).

Across all modality settings, we use the inherited OSTrack objective without an auxiliary BiLift loss:

$$\mathscr{L}_{\text{track}}(\Theta) = \mathscr{L}_{\text{focal}} + 2\mathscr{L}_{\text{GIoU}} + 5\mathscr{L}_{1} \tag{28}$$

Here, $\mathscr{L}_{\text{focal}}$ is the modified CornerNet heatmap loss [24, 25], while $\mathscr{L}_{\text{GIoU}}$ [26] and $\mathscr{L}_{1}$ supervise the decoded target box.

4 Experiments

4.1 Implementation Details

Training. CCLT builds on the ViT-Base variant of OSTrack [23] as its base tracker and is fine-tuned in PyTorch on a single NVIDIA RTX 4090 GPU with a global batch size of 64. Training uses LasHeR [27] for RGB-T, DepthTrack [28] for RGB-D and VisEvent [29] for RGB-E. Template and search crops are 128×128 and 256×256; each epoch contains 60,000 pairs sampled causally within a maximum interval of 200 frames. We train for at most 60 epochs with AdamW [30], weight decay 10^{-4} and a tenfold learning-rate drop at epoch 48. The initial learning rate/batch size is $4 \times 10^{-4}/64$ for RGB-T and $2 \times 10^{-4}/64$ for RGB-D and RGB-E. Unless varied, all low-rank modules use rank 8; IntraLoRA is placed at blocks 2, 4, 6, 8, 10 and 12, and BiLift is at blocks 6 and 10. Each structural cell is evaluated every five epochs; training stops after three consecutive evaluations without a strict IoU gain. For evaluation, we select the checkpoint with the highest validation IoU. Unless otherwise specified, the default random seed is 42; experiments requiring multiple runs use seeds 0 and 1341 in order, following the default run.

Inference. Each modality-specific checkpoint reuses the initial template, performs no online adaptation or template replacement and updates the next search region from the predicted state. Project-controlled forward-only and complete-tracker replay timings are reported separately.

4.2 Evaluation Protocols and Metric Definitions

Training and evaluation routes. We train one task-specific checkpoint for each modality. The RGB–T model is trained on the LasHeR training split and evaluated, without target-benchmark fine-tuning, on the LasHeR test set and RGBT234 [27, 31]. Likewise, the RGB–D model is trained on the DepthTrack training split and evaluated on the DepthTrack test set and VOT-RGBD2022 [28, 32]. After train-only configuration selection, the RGB–E model is retrained on all 500 sequences in the VisEvent training split and evaluated on its disjoint 320-sequence test split [29]. Thus, LasHeR, DepthTrack, and VisEvent provide held-out in-dataset evaluation, whereas RGBT234 and VOT-RGBD2022 probe cross-dataset transfer within RGB–T and RGB–D, respectively.

Computing environment. All experiments use a single NVIDIA GeForce RTX 4090 GPU (24 GB) per training or evaluation job. The implementation runs on PyTorch 2.12.1 with CUDA 13.0; training uses FP32 without automatic mixed precision.

Common box measures. For frame t of sequence s, let $B_{s,t}$ and $G_{s,t}$ denote the predicted and ground-truth boxes, let $c(\cdot)$ denote a box center, and let $(w_{s,t}, h_{s,t})$ be the ground-truth width and height. We use center error, size-normalized center error, and intersection over union (IoU):

$$\begin{aligned} d_{s,t} &= \|c(B_{s,t}) - c(G_{s,t})\|_2 \\ \widetilde{d}_{s,t} &= \sqrt{(\Delta c_x / w_{s,t})^2 + (\Delta c_y / h_{s,t})^2} \\ o_{s,t} &= \frac{|B_{s,t} \cap G_{s,t}|}{|B_{s,t} \cup G_{s,t}|} \end{aligned} \tag{29}$$

For a sequence of N_s frames, define the precision, normalized-precision, and success curves as

$$\begin{aligned} P_s(\delta) &= \frac{1}{N_s} \sum_t \mathbf{1}[d_{s,t} \le \delta] \\ NP_s(\nu) &= \frac{1}{N_s} \sum_t \mathbf{1}[\widetilde{d}_{s,t} \le \nu] \\ S_s(\theta) &= \frac{1}{N_s} \sum_t \mathbf{1}[o_{s,t} > \theta] \end{aligned} \tag{30}$$

Angle brackets $\langle \cdot \rangle_s$ below denote an arithmetic mean over test sequences, giving each sequence equal weight.

LasHeR (RGB–T). The four reported scores are

$$\begin{aligned} PR &= \big\langle P_s(20) \big\rangle_s, \qquad NPR = \big\langle NP_s(0.20) \big\rangle_s \\ NPR\text{-}AUC &= \left\langle \frac{1}{51} \sum_{k=0}^{50} NP_s(k/100) \right\rangle_s \\ SR &= \left\langle \frac{1}{21} \sum_{k=0}^{20} S_s(k/20) \right\rangle_s \end{aligned} \tag{31}$$

PR measures center localization at 20 pixels; NPR reduces sensitivity to image resolution and target size. In addition to the standard LasHeR scores, we report NPR-AUC to summarize normalized localization across thresholds. SR is the area under the IoU-success curve and jointly reflects position and scale accuracy.

RGBT234 (RGB–T). RGBT234 provides visible and thermal annotations that may be spatially misaligned. With $m \in \{V, T\}$ denoting the annotation modality, its protocol uses

$$d_{s,t}^{\min} = \min_m d_{s,t}^{(m)}, \qquad o_{s,t}^{\max} = \max_m o_{s,t}^{(m)} \tag{32}$$

Replacing d and o in Eq. (30) by $d^{\min}$ and $o^{\max}$ gives $P_s^{\min}$ and $S_s^{\max}$, and

$$\begin{aligned} MPR &= \big\langle P_s^{\min}(20) \big\rangle_s \\ MSR &= \left\langle \frac{1}{21} \sum_{k=0}^{20} S_s^{\max}(k/20) \right\rangle_s \end{aligned} \tag{33}$$

MPR measures center accuracy against the closer sensor annotation, whereas MSR measures scale-aware overlap against the better-aligned annotation; this reduces the effect of RGB–thermal registration error on tracker evaluation.

DepthTrack (RGB–D). DepthTrack evaluates long-term tracking with a confidence threshold τ. Let $A_{s,t}(\tau)$ be the predicted box when its confidence is at least τ and the empty set otherwise, $G_{s,t}$ the ground truth, $N_s^p(\tau)$ the number of predicted-present frames, and N_s^g the number of ground-truth-visible frames. Its sequence-balanced precision and recall are

$$\begin{aligned} Pr(\tau) &= \left\langle \frac{1}{N_s^p(\tau)} \sum_{t : A_{s,t}(\tau) \neq \emptyset} \mathrm{IoU}(A_{s,t}(\tau), G_{s,t}) \right\rangle_s \\ Re(\tau) &= \left\langle \frac{1}{N_s^g} \sum_{t : G_{s,t} \neq \emptyset} \mathrm{IoU}(A_{s,t}(\tau), G_{s,t}) \right\rangle_s \\ F(\tau) &= \frac{2\,Pr(\tau)\,Re(\tau)}{Pr(\tau) + Re(\tau)} \end{aligned} \tag{34}$$

At $\tau^* = \arg\max_\tau F(\tau)$, we report $PR = Pr(\tau^*)$, $RE = Re(\tau^*)$, and F-score $= F(\tau^*)$. PR captures localization quality when the tracker declares the target present, RE captures overlap-weighted recovery of visible targets, and F-score balances both.

VOT-RGBD2022 (RGB–D). This benchmark uses the short-term multi-start protocol. For anchored fragment a, let L_a be its length, p_a the successfully tracked prefix before failure, and $o_{a,t}$ its per-frame IoU. If A_s is the set of fragments from sequence s and T_s that sequence's length, then

$$\begin{aligned} Acc &= \frac{\sum_a \sum_{t=1}^{p_a} o_{a,t}}{\sum_a p_a} \\ Robs = \frac{\sum_{a\in A_s} p_a}{\sum_{a\in A_s} L_a}, \qquad Rob &= \frac{\sum_s T_s Robs}{\sum_s T_s} \end{aligned} \tag{35}$$

Let $\phi(\ell)$ be the official anchor-weighted expected mean overlap for a fragment of length ℓ, with post-failure overlap set to zero. The expected average overlap is

$$EAO = \frac{1}{755-115+1} \sum_{\ell=115}^{755} \phi(\ell) \tag{36}$$

Accuracy measures localization before failure, robustness measures how much of the anchored fragments remains successfully tracked, and EAO combines both over representative sequence lengths.

VisEvent (RGB–E). VisEvent follows the PR/SR definitions in Eq. (30) with its target-absence labels. Let $a_{s,t} = 1$ indicate a present target and zero otherwise. Its scores are

$$\begin{aligned} PR &= \left\langle \frac{1}{N_s} \sum_t a_{s,t} \mathbf{1}[d_{s,t} \le 20] \right\rangle_s \\ SR &= \left\langle \frac{1}{21} \sum_{k=0}^{20} \frac{1}{N_s} \sum_t a_{s,t} \mathbf{1}[o_{s,t} > k/20] \right\rangle_s \end{aligned} \tag{37}$$

Absent frames do not contribute localization successes, while the complete sequence length normalizes the curves. PR therefore measures 20-pixel center accuracy and SR summarizes scale-aware overlap. All reported metrics are higher-is-better.

The parameter-matched BiLift–Parallel-LR comparison is restricted to RGB–T and pre-specifies LasHeR SR as its primary metric. The two variants use the same frozen backbone, trainable-parameter budget, data split, optimizer, schedule, augmentation, and validation-only checkpoint-selection rule; only the corrected-state versus parallel update topology differs. We report training seeds 0 and 1 separately and in an equal-weight summary.

For each metric, uncertainty is estimated with 10000 paired sequence-level bootstrap resamples within each run. The same resampled sequences are used for both methods before computing their difference, and the two runs receive equal weight. The resulting 95% intervals quantify test-sequence sampling variation conditional on the two fixed checkpoints; they do not estimate variation over training seeds.

4.3 Comparison with State-of-the-Art Trackers

Datasets and metrics. We evaluate separate modality checkpoints on LasHeR/RGBT234 (RGB-T) [27, 31], DepthTrack/VOT-RGBD2022 (RGB-D) [28, 32] and VisEvent (RGB-E) [29]. We report PR/NPR/SR and MPR/MSR for RGB-T, PR/RE/F-score and EAO/Acc./Rob. for RGB-D and PR/SR for RGB-E; all scores are percentages and higher is better. LasHeR SR is the controlled primary because IoU-success AUC reflects localization and scale; Sect. 4.2 provides formal metric definitions.

Overall comparison. Table 1 compares CCLT with 9 fully fine-tuned (FFT) trackers and 13 prior parameter-efficient or training-free trackers across five RGB-X benchmarks. To reflect their different adaptation budgets, the best and second-best results are marked separately within the FFT and PEFT/TF blocks.

Overall results. Within the PEFT/TF block, CCLT records the highest LasHeR PR, NPR and SR. It ranks second on both RGBT234 metrics and all three DepthTrack metrics, and obtains the highest VOT-RGBD2022 EAO while tying the highest accuracy. On VisEvent, CCLT ranks second in PR among PEFT/TF methods.

PEFT efficiency. CCLT updates 196.6K parameters, only 0.212% of its 92.7M-parameter tracker. This is the smallest reported trainable budget among the listed learned PEFT trackers with available counts. Under the same project-controlled network-forward-only protocol, CCLT runs at 106.8 FPS, the highest inference speed among all methods in Table 1 under this controlled protocol.

Prediction-level comparison. Across the displayed RGB-T, RGB-D and RGB-E cases in Fig. 4, competing trackers drift or confuse the target, whereas CCLT remains closer to the ground truth. This cross-modal consistency accords with BiLift's progressive residual correction.

Table 1 Performance comparison across RGB-T, RGB-D and RGB-E benchmarks. All tracking scores are percentages and "–" denotes an unreported result. L.P. denotes learnable parameters in millions. Bold and underlined values mark the best and second-best available values, computed separately within the FFT and PEFT/TF blocks; tied values receive the same mark

			LasHeR			RGBT234		DepthTrack			VOT-RGBD2022			VisEvent		L.P.	FPS
Type	Method	Source	PR	NPR	SR	MPR	MSR	PR	RE	F	EAO	Acc.	Rob.	PR	SR	(M)↓	
FFT	TBSI (2023)	CVPR[18]	69.2	65.7	55.6	87.1	63.7	–	–	–	–	–	–	–	–	203	36.2
	IIMF (2024)	MM[19]	**72.4**	**68.4**	**58.1**	86.8	64.4	–	–	–	–	–	–	–	–	182	–
	CAFormer (2025)	AAAI[20]	70.0	66.1	55.6	**88.3**	**66.4**	–	–	–	–	–	–	–	–	–	83.6
	DeT (2021)	CVPR	–	–	–	–	–	56.0	50.6	53.2	65.7	76.0	84.5	–	–	–	–
	TABBTrack (2025)	PR[33]	–	–	–	–	–	62.2	61.5	61.8	72.2	82.1	87.4	–	–	–	27.0
	PHPTrack (2025b)	TCSVT[34]	–	–	–	–	–	–	–	–	–	–	–	77.8	61.1	93.7	21.5
	CMDTrack (2025)	TPAMI[35]	68.8	–	56.6	85.9	61.8	59.1	60.7	59.8	–	–	–	75.8	61.3	**31**	67.0
	MamTrack (2025)	CVPR[36]	67.4	–	54.2	84.4	62.4	–	–	–	–	–	–	**79.2**	**61.6**	–	18.1
	SMSTrack (2025)	ICCV[37]	70.3	–	56.0	86.9	64.5	**64.1**	**63.1**	**63.6**	**74.8**	**82.2**	**89.7**	76.3	60.4	98.2	24.1
PEFT/TF	ProTrack (2022)	MM[16]	53.8	49.8	42.0	79.5	59.9	58.3	57.3	57.8	65.1	80.1	80.2	63.2	47.1	–	–
	ViPT (2023)	CVPR[1]	65.1	61.6	52.5	83.5	61.7	59.2	59.6	59.4	72.1	81.5	87.1	75.8	59.2	0.8	39.2
	VADT (2024)	ICASSP[38]	–	–	–	–	–	60.3	60.6	61.0	72.1	81.6	87.3	–	–	–	–
	TATrack (2024a)	AAAI[39]	70.2	66.7	56.1	87.2	64.4	–	–	–	–	–	–	–	–	–	26.1
	BAT (2024)	AAAI[6]	70.2	66.4	56.3	86.8	64.1	–	–	–	–	–	–	–	–	0.3	43.7
	eMoET (2025)	RAL[40]	–	–	–	–	–	–	–	–	–	–	–	76.4	61.3	8.4	–
	UnTrack (2024)	CVPR[41]	64.6	60.0	51.3	84.2	62.5	61.1	60.8	61.0	72.1	82.0	86.9	75.5	58.9	6.6	25.6
	OneTracker (2024)	CVPR[4]	67.2	–	53.8	85.7	64.2	60.7	60.4	60.9	72.7	81.9	87.2	76.7	60.8	2.8	–
	SDSTrack (2024)	CVPR[2]	66.5	62.6	53.1	84.8	62.5	61.9	60.9	61.4	72.8	81.2	88.3	76.7	59.7	14.8	20.8
	SNNPTrack (2025)	ICASSP[42]	–	–	–	–	–	–	–	–	–	–	–	76.9	59.8	2.1	–
	M3Track (2025)	SPL[43]	65.8	–	52.5	84.5	63.0	56.4	58.5	57.4	–	–	–	76.7	59.6	1.7	31.1
	XTrack (2025)	ICCV[12]	69.1	65.5	55.7	87.4	64.9	61.8	62.0	61.5	74.0	**82.1**	**88.8**	**77.5**	60.9	5.4	10.3
	SEATrack (2026)	CVPR[5]	71.6	67.5	57.3	**87.8**	63.9	**62.9**	**63.5**	**63.2**	73.6	**82.1**	88.4	77.1	60.3	0.6	63.5
	CCLT (ours)	Ours	**71.8**	**67.9**	**57.6**	87.5	64.5	62.4	62.9	62.3	**74.1**	**82.1**	88.7	77.2	60.6	**0.2**	106.8

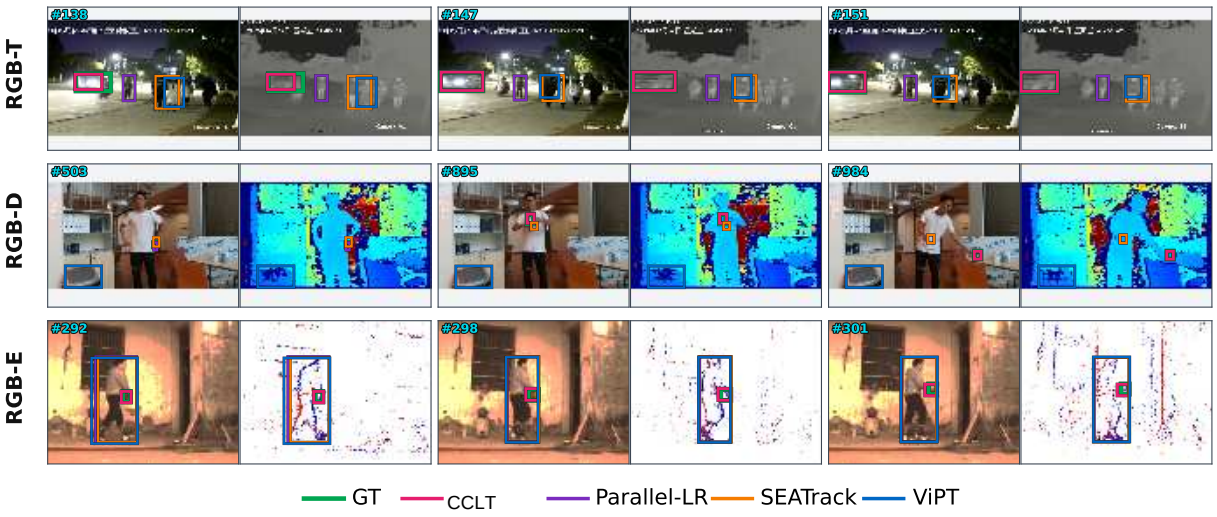


Figure 4 Prediction-level comparison of CCLT with representative PEFT-based multimodal trackers across RGB-T, RGB-D and RGB-E tasks

4.4 Ablation Studies

Key component analysis. Table 2 shows that S0-to-S1 improves all eight metrics, by 15.35–20.13 points on LasHeR, 5.74–8.94 on DepthTrack and 7.49–8.02 on VisEvent. Although not parameter matched, these gains indicate that IntraLoRA effectively adapts frozen within-stream representations. S2-to-S3 is mixed (-0.74 to $+0.54$ points), whereas S4-to-S5 improves every metric: $+2.20$–3.10 on LasHeR, $+0.22$–0.82 on DepthTrack and $+0.67$–1.22 on VisEvent. This matched contrast shows that IntraLoRA prepares adapted stream states on which BiLift can progressively refine earlier corrections; their complementarity explains the cross-task gains.

Table 2 Component ablation on LasHeR, DepthTrack and VisEvent (percent). ✓/× denote enabled/disabled components; "Corrected" marks whether the second branch reads the corrected state (BiLift) or the incoming state (Parallel-LR)

		Adaptation only		Exchange, no LoRA		Exchange + LoRA	
Group	Metric	S0	S1	S2	S3	S4	S5
Setup	IntraLoRA	×	✓	×	×	✓	✓
	Exchange	×	×	✓	✓	✓	✓
	Corrected	–	–	×	✓	×	✓
	L.P. (K)	0.0	147.5	49.2	49.2	196.6	196.6
LasHeR	PR	48.17	68.3	67.78	67.06	68.7	**71.8**
	NPR	44.75	64.3	64.04	63.49	65.2	**67.9**
	SR	39.35	54.7	54.22	53.88	55.4	**57.6**
DepthTrack	PR	52.26	61.2	60.98	61.3	61.58	**62.4**
	RE	53.97	59.71	60.44	59.7	62.4	**62.9**
	F	53.10	60.45	60.71	60.49	62.08	**62.3**
VisEvent	PR	67.92	75.94	75.13	75.42	76.53	**77.2**
	SR/AUC	52.10	59.59	58.63	59.17	59.38	**60.6**

Effect of corrected-state dependency. Table 2 motivates a three-seed, parameter-matched test that changes only the second branch's source state. At the same 196.6K-parameter budget, CCLT gains $+1.980$ LasHeR SR points, with a 95% interval of $[0.360, 2.690]$. With all other factors controlled, the positive interval attributes the gain to corrected-state reading: the later branch refines the earlier update, forming a state-progressive relay without extra parameters.

Fixed-weight path interventions. With weights, data and the evaluator fixed, Table 3 shows that every path intervention reduces LasHeR SR. Disabling both sites hurts most ($-13.19/-3.87$ points), indicating cross-depth complementarity. Removing either direction also hurts, especially RGB$\to X$ (-9.33 versus -7.13; -3.47 versus -2.73). Stale sources cost another 0.33–1.16 points, confirming that fresh-state reuse drives progressive cross-modal refinement.

Table 3 Fixed-weight path interventions on LasHeR. Values denote SR changes (%) from the clean run. "Site off" disables BiLift at the listed blocks, "Direction off" removes one directional correction and "Stale source" forces the second branch to read the incoming state

	Site off			Direction off		Stale source		
Config.	B6	B10	Both	$X\to$RGB	RGB$\to X$	B6	B10	Both
S3	-8.36	-5.19	-13.19	-7.13	-9.33	-1.07	-0.66	-1.16
S5	-1.70	-2.77	-3.87	-2.73	-3.47	-1.07	-0.33	-0.58

Correction-energy analysis. Figure 5 shows subunit update-to-state ratios, low-dimensional correction energy, target-focused endpoints and modality-specific directional profiles on RGB-T. The RGB-D and RGB-E counterparts in Sect. 4.5 show the same trends. Together with Tables 2 and 3, these results indicate compact, target-focused and modality-adaptive BiLift corrections.

Direction-order sensitivity. Reversing the initiating order at $K = 2$, $r = 8$ reduces SR by $1.047/0.898$ points on LasHeR/RGBT234 with fixed weights and by $0.597/0.673$ after two-seed retraining; all four 95% paired-bootstrap intervals are below zero. The consistent losses show that BiLift's depth-aware alternating leadership better exploits complementary directional corrections.

Missing-modality robustness. Table 4 shows consistent S5 gains when either modality is missing. Without RGB/thermal, all four tracking metrics improve by 5.03–6.46/2.88–3.40 points, while SR retention rises by

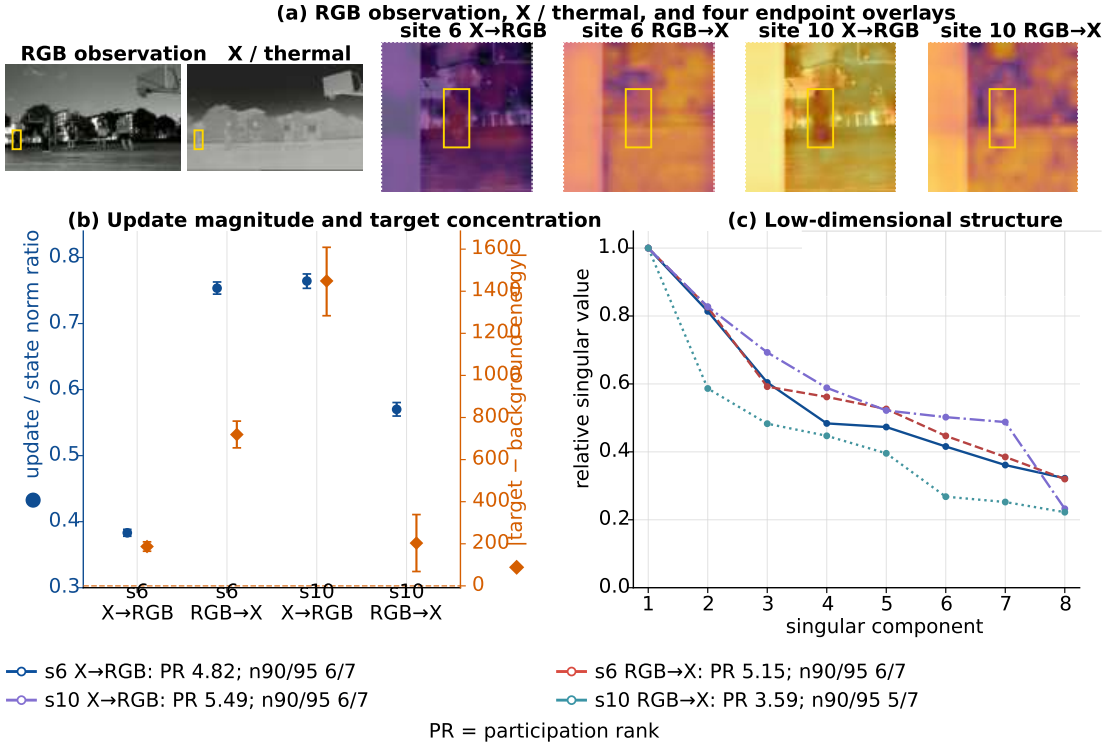


Figure 5 BiLift diagnostics on LasHeR (RGB-T): subunit update-to-state ratios, low-dimensional correction energy, target-focused endpoints and modality-specific directional profiles

7.75/2.56 points. This indicates that corrected-state reading better preserves useful cues from the remaining stream; the larger RGB-missing loss reveals stronger RGB dependence. The results support relative missing-modality robustness.

Table 4 LasHeR missing-modality robustness. $\Delta = \text{S5} - \text{S4}$; SR loss is clean-minus-missing; Ret. is missing/clean SR (%), with ΔRet. in points

Missing	Metric	S4	S5	Δ
RGB	PR ↑	13.43	**19.89**	+6.46
	NPR ↑	13.11	**19.16**	+6.05
	NPR-AUC ↑	12.77	**18.36**	+5.59
	SR ↑	14.47	**19.51**	+5.03
	SR loss ↓	40.93	**38.09**	−2.83
	SR Ret. (%) ↑	26.12	**33.87**	+7.75
Thermal	PR ↑	43.14	**46.54**	+3.40
	NPR ↑	40.47	**43.59**	+3.13
	NPR-AUC ↑	37.02	**40.08**	+3.06
	SR ↑	35.45	**38.33**	+2.88
	SR loss ↓	19.95	**19.27**	−0.68
	SR Ret. (%) ↑	63.99	**66.55**	+2.56

Attribute-level analysis. Figure 6 shows S5 improves PR and SR on 18 of 19 attributes, with the largest gains on HI (+6.0/ + 4.0 PR/SR), LR (+6.0/ + 4.0) and BC (+5.3/ + 4.6). This broad advantage suggests that BiLift's corrected-state relay better exploits complementary cues under degraded or ambiguous appearance; the small mixed changes on OV and AIV show that the benefit is not uniform.

Position. Table 5(a) shows that $\{6, 10\}$ leads four metrics and ties the best MSR (64.5), while $\{4, 12\}$ trails by 0.8–1.2 points on the others. Differences among equally spaced pairs indicate that BiLift gains from complementary corrections at specific representation depths, not spacing alone.

Site count. Table 5(b) shows that $K = 2$ is best throughout. Increasing to $K = 12$ lowers SR/MSR by 2.5/1.2 points, while parameters reach 294.9K at $K = 6$. Thus, sparse relays suffice; denser exchange adds cost without gain. Because placements also vary across rows, this remains configuration-level rather than isolated count evidence.

Rank. Table 5(c) shows that $r = 8$ leads all five accuracy metrics. Rank 4 is cheaper and faster (172.0K, 113.4

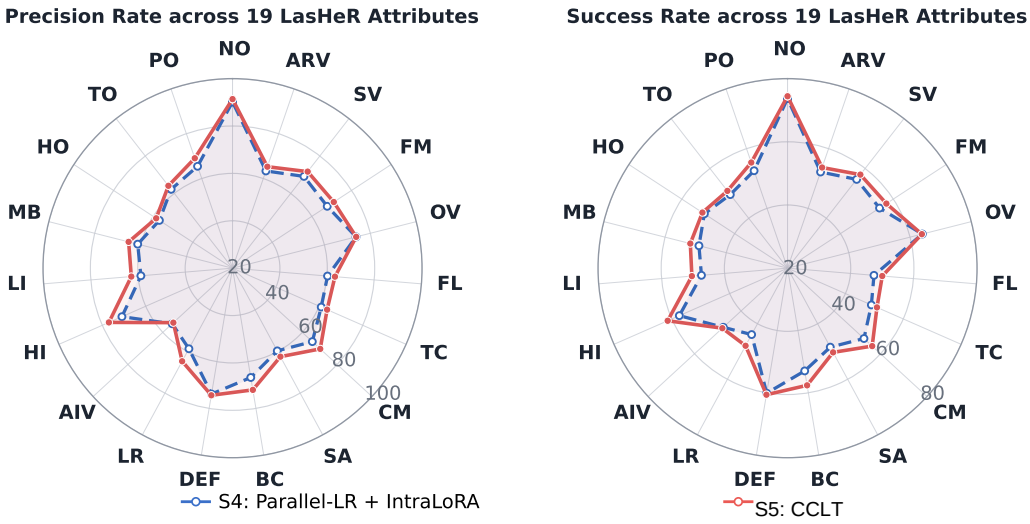


Figure 6 S4 versus S5 across 19 LasHeR attributes. S5 improves PR and SR on 18 of 19 attributes, with the largest gains on HI, LR and BC

Table 5 BiLift design choices on LasHeR/RGBT234 (%; one decimal). Defaults are sites $\{6,10\}$ and $r = 8$; L.P. is in thousands and "all" denotes 12 blocks. Bold/underlined values mark the best/second best; ties share marks

(a) Position ($r = 8$, two sites)

Sites	PR ↑	NPR ↑	SR ↑	MPR ↑	MSR ↑
$\{2,6\}$	68.7	65.0	55.2	86.2	63.7
$\{4,8\}$	69.7	65.7	55.8	86.3	63.7
$\{6,10\}$	**71.8**	**67.9**	**57.6**	**87.5**	**64.5**
$\{8,12\}$	70.2	66.4	56.5	86.4	63.9
$\{7,9\}$	69.9	65.8	55.9	86.6	63.7
$\{4,12\}$	70.6	<u>66.8</u>	<u>56.8</u>	86.7	**64.5**

(b) Site Count ($r = 8$)

K	Sites	PR ↑	NPR ↑	SR ↑	MPR ↑	MSR ↑	L.P. ↓
2	$\{6,10\}$	**71.8**	**67.9**	**57.6**	**87.5**	**64.5**	**196.6**
3	$\{6,10,12\}$	70.2	66.2	56.3	85.4	63.0	221.2
4	$\{6,8,10,12\}$	68.6	65.1	55.3	86.9	64.0	245.8
6	even blocks	68.8	65.1	55.3	85.8	63.5	294.9
12	all	68.5	64.9	55.1	85.6	63.3	–

(c) Rank (sites $\{6,10\}$)

r	PR ↑	NPR ↑	SR ↑	MPR ↑	MSR ↑	L.P. ↓	FPS ↑
4	<u>70.1</u>	<u>66.6</u>	<u>56.4</u>	<u>86.5</u>	<u>64.1</u>	**172.0**	**113.4**
8	**71.8**	**67.9**	**57.6**	**87.5**	**64.5**	196.6	106.8
16	69.6	65.8	55.8	86.0	62.9	245.8	103.7

FPS) but less accurate; rank 16 costs 245.8K, slows to 103.7 FPS and lowers SR/MSR by 1.8/1.6 points. Thus, rank 8 best balances correction capacity, tracking accuracy and inference efficiency within the tested design space.

4.5 Cross-Task Correction-Energy Diagnostics

Figures 7, 8 and 9 present this correction-energy diagnostic at full width on LasHeR (RGB-T), DepthTrack (RGB-D), and VisEvent (RGB-E), respectively. Across the three tasks, BiLift produces subunit update-to-state ratios and concentrates 90% of its correction energy in three to six singular components. The endpoint overlays and target–background energy differences further show target-focused, modality-dependent directional corrections. These cross-task trends support compact and adaptive behavior.

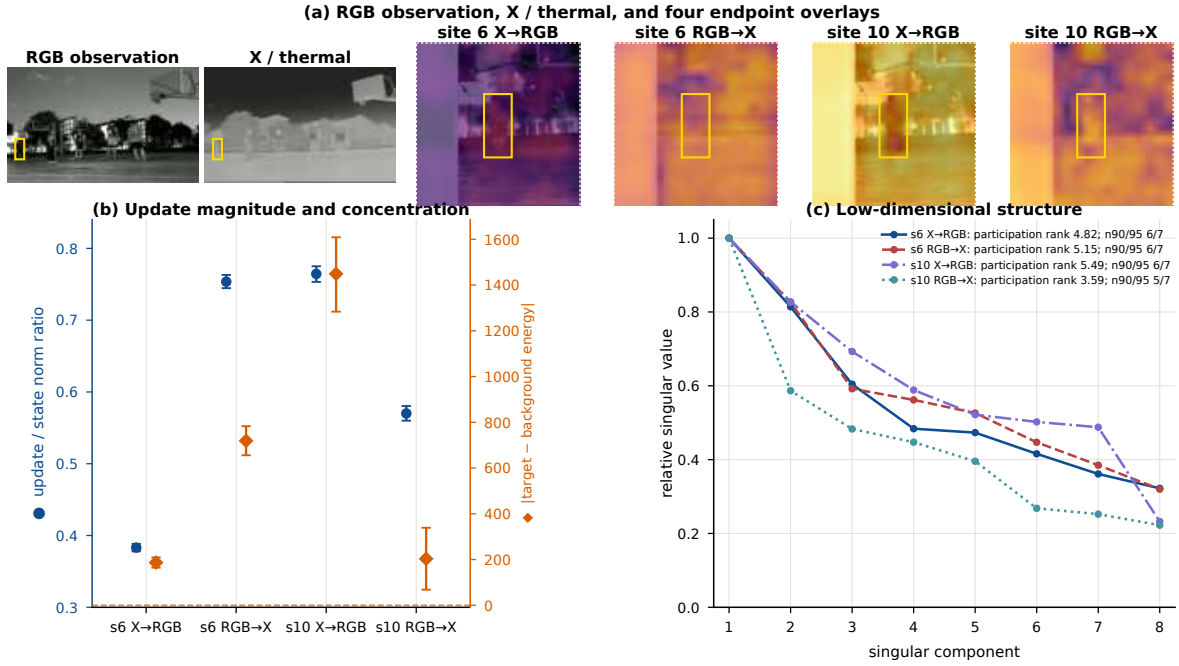


Figure 7 BiLift correction-energy diagnostics on LasHeR (RGB-T): endpoint overlays, update magnitude and target concentration, and the singular-value structure of the four directional corrections

4.6 Qualitative Case Studies

Figures 10–12 show the two highest-ranked clear-win sequences from each frozen candidate pack. Each sequence contains four temporally separated RGB–X frame pairs and the predictions of CCLT, parameter-matched Parallel-LR, SEATrack, and ViPT. The source sheets report the selection metric and per-frame IoU, and exclude initialization from candidate scoring. Here, clear-win denotes the candidate-pack selection role; it does not mean that CCLT wins every frame or the complete sequence.

The three modality families reveal the same qualitative separation in different sensing conditions. In RGB–T, the thermal stream exposes foreground structure when visible contrast is weak, but the competing predictions often lock onto a larger bright or warm distractor. In RGB–E, the event view contains sparse and fragmented motion evidence; CCLT preserves the target's compact extent instead of absorbing the surrounding person or vehicle. In RGB–D, depth discontinuities help distinguish a manipulated object from the nearby carrier, and CCLT follows that target-scale region through the interaction.

Most importantly, the separation from parameter-matched Parallel-LR is visible across all three modality types: CCLT remains aligned after the parallel control has drifted or expanded. This behavior is consistent with BiLift's corrected-state relay, in which the second directional correction reads the state already updated by the first and can reconcile complementary evidence within the same interaction site without adding trainable parameters relative to Parallel-LR.

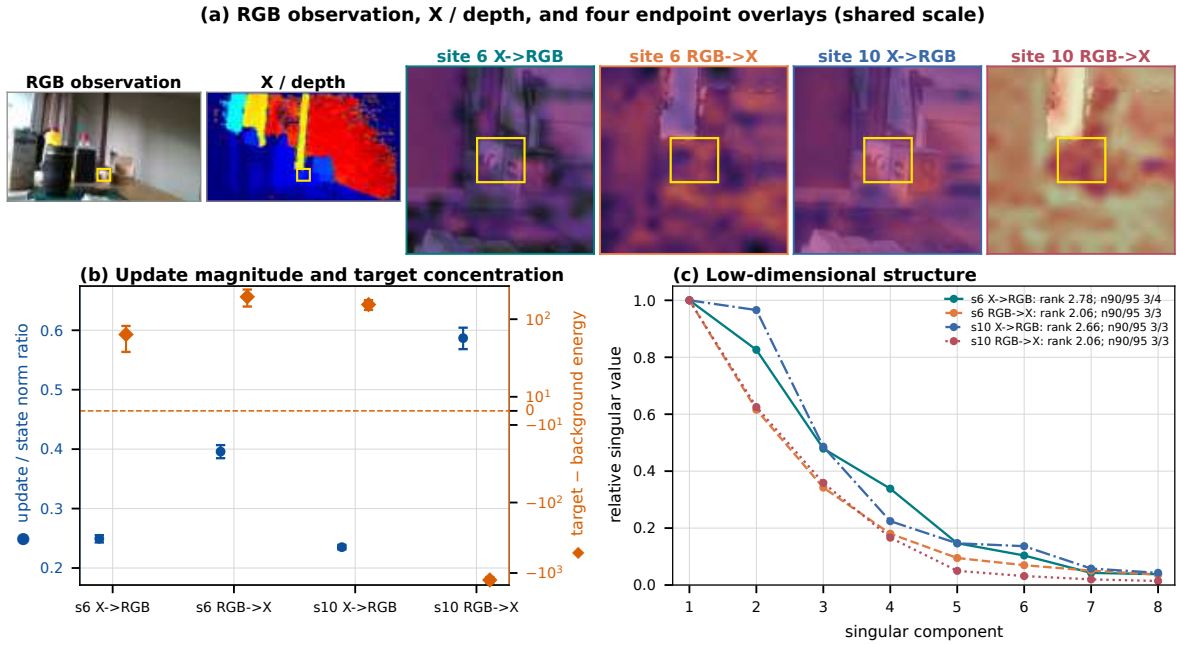


Figure 8 BiLift correction-energy diagnostics on DepthTrack (RGB-D): endpoint overlays, update magnitude and target concentration, and the singular-value structure of the four directional corrections

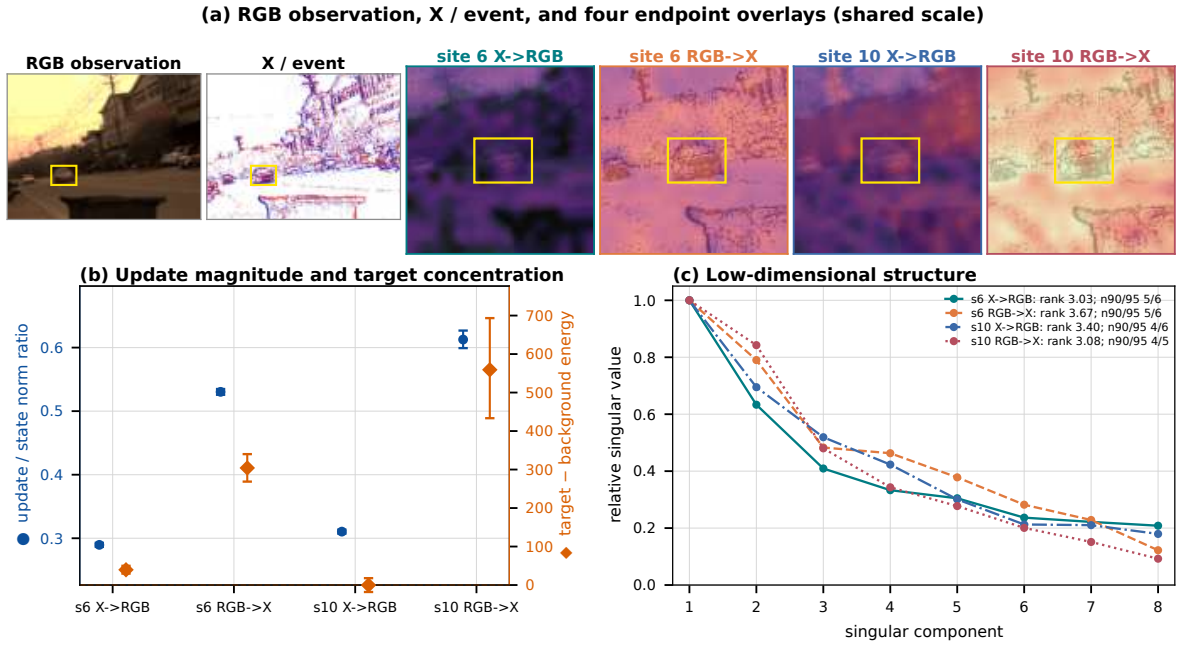


Figure 9 BiLift correction-energy diagnostics on VisEvent (RGB-E): endpoint overlays, update magnitude and target concentration, and the singular-value structure of the four directional corrections

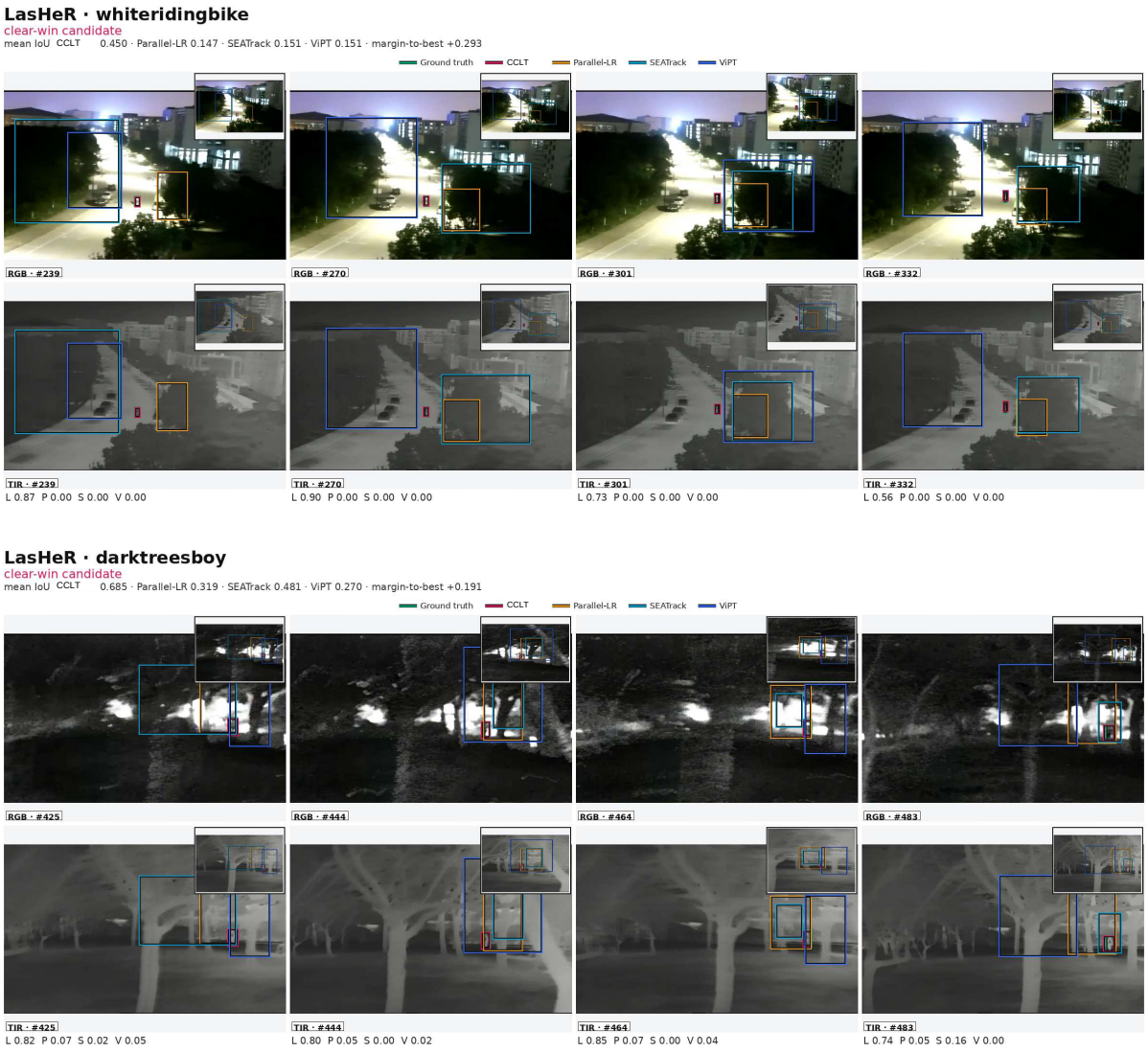


Figure 10 Candidate-pack clear-win comparisons on LasHeR (RGB–T). The upper and lower source sheets are the first two entries in the recorded LasHeR clear-win ranking. Each sheet shows four selected moments, with RGB above TIR. Boxes denote ground truth, CCLT, parameter-matched Parallel-LR, SEATrack, and ViPT, using the embedded legend; per-frame IoUs appear below each column. In both sequences, CCLT retains a compact target-aligned box under low illumination, small target scale, and strong background distractors, while the comparators drift toward larger high-contrast regions

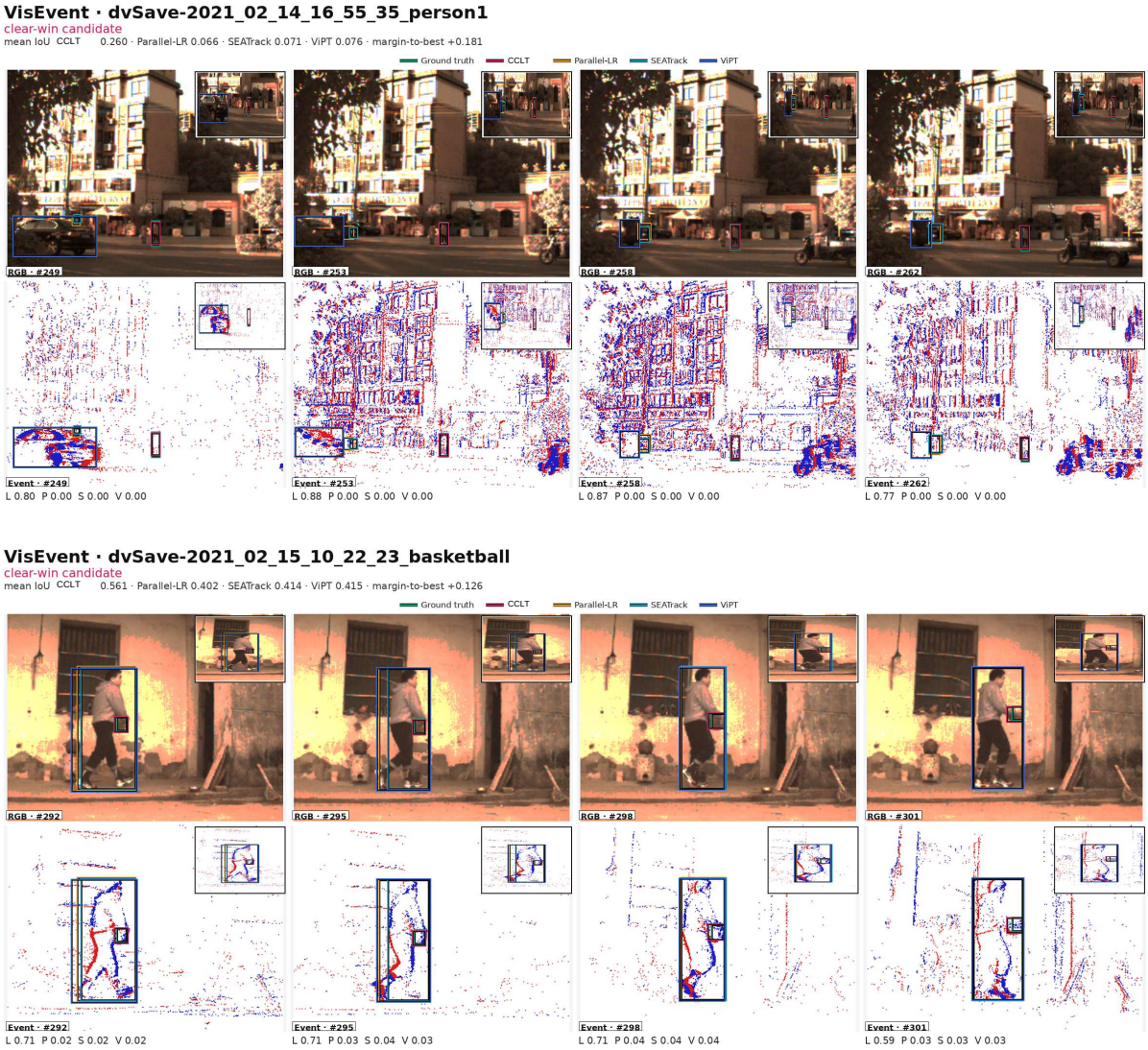


Figure 11 Candidate-pack clear-win comparisons on VisEvent (RGB–E). The upper and lower source sheets are the first two entries in the recorded VisEvent clear-win ranking. Each sheet shows RGB above the event representation. The top sequence contains a small pedestrian amid vehicle and edge clutter; the bottom sequence tracks a basketball near a moving person. CCLT remains close to the compact ground-truth extent, whereas the comparison methods often attach to a larger object or person region. Box and IoU conventions follow Fig. 10

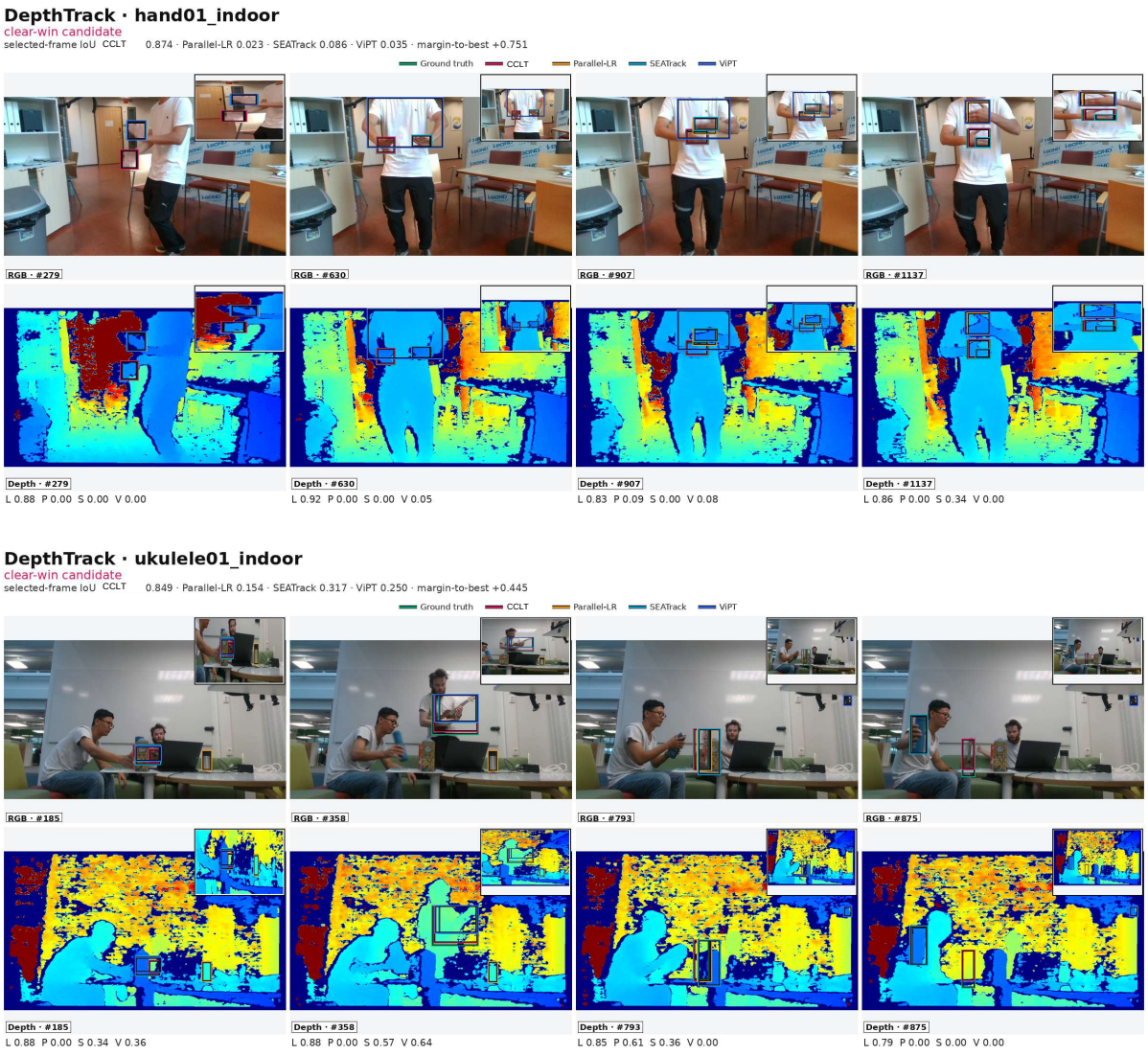


Figure 12 Candidate-pack clear-win comparisons on DepthTrack (RGB–D). The upper and lower source sheets are the first two entries in the recorded DepthTrack clear-win ranking. Each sheet shows RGB above pseudocolored depth. Depth is clipped at $\min(3 \times \text{median}, 10000)$, min–max normalized, and rendered with JET for display. These scenes combine hand-held targets, nearby people, scale change, and partial occlusion. CCLT follows the target-sized RGB/depth region across the four selected moments, while the comparison boxes more often expand to the carrier or move to a nearby object. The header reports selected-frame IoU rather than complete-sequence mean IoU; box conventions follow Fig. 10

5 Conclusion

We revisit state dependency in PEFT RGB-X interaction with CCLT. At two sparse depths, BiLift turns parallel updates into a corrective relay over freshly corrected states, while modality-shared IntraLoRA adapts both streams on a frozen backbone. Across five benchmarks, controlled experiments validate fresh-state reuse and show a favorable performance–efficiency trade-off with only 196.6K trainable parameters (0.212% of the tracker).

Limitations. CCLT remains sensitive to missing modalities: removing RGB/thermal lowers LasHeR SR by 38.09/19.27 points (Table 4). Future work will explore missing-modality augmentation, reliability-aware routing, and sparse correction for partially observed or spatially misaligned inputs in practical scenarios.

Declarations

Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 62605035).

Conflict of interest

The authors have no relevant financial or non-financial interests to disclose.

Ethics approval

Not applicable.

Consent to participate and consent for publication

Not applicable.

Data availability

The experiments in this study use five publicly available benchmark datasets: LasHeR (`https://github.com/BUGPLEASEOUT/LasHeR`), RGBT234 (`https://chenglongli.cn/Datasets-and-benchmark-code/`), DepthTrack (`https://github.com/xiaozai/DeT`), VOT-RGBD2022 (`https://www.votchallenge.net/vot2022/dataset.html`), and VisEvent (`https://github.com/wangxiao5791509/VisEvent_SOT_Benchmark`). The datasets and annotations are available through their respective official project pages and download instructions; VOT-RGBD2022 can be obtained through the VOT toolkit. Access and use are subject to the terms specified by the dataset providers. The streamlined source code, the RGB-T model checkpoint at epoch 60, and checksum files are publicly available in the Zenodo archive at `https://doi.org/10.5281/zenodo.23483732` .

Author contributions

Xingyu Tan: Conceptualization, Methodology, Software, Writing – original draft. Chun Liu: Funding acquisition, Project administration, Writing – review & editing. Qing Song: Funding acquisition, Project administration, Writing – review & editing. Mengjie Hu: Funding acquisition, Project administration, Writing – review & editing, Supervision.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62605035).

Use of AI tools

AI tools were used solely for language polishing, including grammar, spelling, punctuation, and tone.

References

[1] J. Zhu, S. Lai, X. Chen, D. Wang, and H. Lu. Visual prompt multi-modal tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9516–9526, 2023. DOI: https://doi.org/10.1109/CVPR52729.2023.00918.

[2] X. Hou, J. Xing, Y. Qian, Y. Guo, S. Xin, J. Chen, K. Tang, M. Wang, Z. Jiang, L. Liu, and Y. Liu. SDSTrack: self-distillation symmetric adapter learning for multi-modal visual object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26551–26561, 2024. DOI: https://doi.org/10.1109/CVPR52733.2024.02507.

[3] H. Wang, T. Xu, Z. Tang, X.-J. Wu, and J. Kittler. Multi-modal adapter for RGB-T tracking. *Information Fusion*, 118:102940, 2025. DOI: https://doi.org/10.1016/j.inffus.2025.102940.

[4] L. Hong, S. Yan, R. Zhang, W. Li, X. Zhou, P. Guo, K. Jiang, Y. Chen, J. Li, Z. Chen, and W. Zhang. OneTracker: unifying visual object tracking with foundation models and efficient tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19079–19091, 2024. DOI: https://doi.org/10.1109/CVPR52733.2024.01805.

[5] J. Su, Z. Xue, S. Zhang, K. Chen, W. Hu, and Z. Zhang. SEATrack: simple, efficient, and adaptive multimodal tracker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28679–28689, 2026. DOI: https://doi.org/10.1109/CVPR52734.2026.02687.

[6] B. Cao, J. Guo, P. Zhu, and Q. Hu. Bi-directional adapter for multimodal tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 927–935, 2024. DOI: https://doi.org/10.1609/aaai.v38i2.27852.

[7] Q. Hu, Y. Li, and W. Yu. Exploiting multimodal prompt learning and distillation for RGB-T tracking. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, pages 469–477, 2025. DOI: https://doi.org/10.1145/3731715.3733332.

[8] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727, 2022. DOI: https://doi.org/10.1007/978-3-031-19827-4_41.

[9] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 2790–2799, 2019. DOI: https://doi.org/10.48550/arXiv.1902.00751.

[10] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. DOI: https://doi.org/10.48550/arXiv.2106.09685.

[11] L. Lin, H. Fan, Z. Zhang, Y. Wang, Y. Xu, and H. Ling. Tracking meets LoRA: faster training, larger model, stronger performance. In *European Conference on Computer Vision*, pages 300–318, 2024. DOI: https://doi.org/10.1007/978-3-031-73232-4_17.

[12] Y. Tan, Z. Wu, Y. Fu, Z. Zhou, G. Sun, E. Zamfir, C. Ma, D. Paudel, L. Van Gool, and R. Timofte. XTrack: multimodal training boosts RGB-X video object trackers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5734–5744, 2025. DOI: https://doi.org/10.1109/ICCV51701.2025.00543.

[13] Z. Shao, Y. Hu, J. Liu, B. Fan, and H. Liu. Group orthogonal low-rank adaptation for RGB-T tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 8887–8895, 2026. DOI: https://doi.org/10.1609/aaai.v40i11.37843.

[14] Y. Zhu, J. Li, C. Li, J. Wang, C. Gu, and J. Tang. Sparse-dense mixture of experts adapter for multi-modal tracking. arXiv:2603.13719, 2026. DOI: https://doi.org/10.48550/arXiv.2603.13719.

[15] H. Wang, T. Xu, Z. Tang, X.-J. Wu, and J. Kittler. Learning progressive adaptation for multi-modal tracking. arXiv:2603.21100, 2026. DOI: https://doi.org/10.48550/arXiv.2603.21100.

[16] J. Yang, Z. Li, F. Zheng, A. Leonardis, and J. Song. Prompting for multi-modal tracking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3492–3500, 2022. DOI: https://doi.org/10.1145/3503161.3547851.

[17] Z. Ou, D. Zhang, G. Ying, and Z. Zheng. UBPT: unidirectional and bidirectional prompts for RGBD tracking. *IEEE Sensors Journal*, 24(22):37503–37513, 2024. DOI: https://doi.org/10.1109/JSEN.2024.3470852.

[18] T. Hui, Z. Xun, F. Peng, J. Huang, X. Wei, X. Wei, J. Dai, J. Han, and S. Liu. Bridging search region interaction with template for RGB-T tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13630–13639, 2023. DOI: https://doi.org/10.1109/CVPR52729.2023.01310.

[19] L. Chen, Y. Huang, H. Li, Z. Zhou, and Z. He. Simplifying cross-modal interaction via modality-shared features for RGBT tracking. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1573–1582, 2024. DOI: https://doi.org/10.1145/3664647.3681564.

[20] Y. Xiao, J. Zhao, A. Lu, C. Li, B. Yin, Y. Lin, and C. Liu. Cross-modulated attention transformer for RGBT tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8682–8690, 2025. DOI: https://doi.org/10.1609/aaai.v39i8.32938.

[21] A.N. Gomez, M. Ren, R. Urtasun, and R.B. Grosse. The reversible residual network: backpropagation without storing activations. In *Advances in Neural Information Processing Systems*, volume 30, pages 2214–2224, 2017. https://proceedings.neurips.cc/paper/2017/hash/f9be311e65d81a9ad8150a60844bb94c-Abstract.html.

[22] H. Zhao, M. Li, Q. Hu, and X. Guo. Reversible decoupling network for single image reflection removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26430–26439, 2025. DOI: https://doi.org/10.1109/CVPR52734.2025.02461.

[23] B. Ye, H. Chang, B. Ma, S. Shan, and X. Chen. Joint feature learning and relation modeling for tracking: a one-stream framework. In *European Conference on Computer Vision*, pages 341–357, 2022. DOI: https://doi.org/10.1007/978-3-031-20047-2_20.

[24] H. Law and J. Deng. CornerNet: detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision*, pages 734–750, 2018. DOI: https://doi.org/10.1007/978-3-030-01264-9_48.

[25] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Doll'ar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017. DOI: https://doi.org/10.1109/ICCV.2017.324.

[26] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese. 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*, pages 658–666, 2019. DOI: https://doi.org/10.1109/CVPR.2019.00075.

[27] C. Li, W. Xue, Y. Jia, Z. Qu, B. Luo, J. Tang, and D. Sun. LasHeR: a large-scale high-diversity benchmark for RGBT tracking. *IEEE Transactions on Image Processing*, 31:392–404, 2022. DOI: https://doi.org/10.1109/TIP.2021.3130533.

[28] S. Yan, J. Yang, J. K"apyl"a, F. Zheng, A. Leonardis, and J.-K. K"am"ar"ainen. DepthTrack: unveiling the power of RGBD tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10725–10733, 2021. DOI: https://doi.org/10.1109/ICCV48922.2021.01055.

[29] X. Wang, J. Li, L. Zhu, Z. Zhang, Z. Chen, X. Li, Y. Wang, Y. Tian, and F. Wu. VisEvent: reliable object tracking via collaboration of frame and event flows. *IEEE Transactions on Cybernetics*, 54(3):1997–2010, 2024. DOI: https://doi.org/10.1109/TCYB.2023.3318601.

[30] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. DOI: https://doi.org/10.48550/arXiv.1711.05101.

[31] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang. RGB-T object tracking: benchmark and baseline. *Pattern Recognition*, 96:106977, 2019. DOI: https://doi.org/10.1016/j.patcog.2019.106977.

[32] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, J.-K. K"am"ar"ainen, H.J. Chang, M. Danelljan, L.C. Zajc, et al. The tenth visual object tracking VOT2022 challenge results. In *Computer Vision – ECCV 2022 Workshops*, pages 431–460. Springer, 2023. DOI: https://doi.org/10.1007/978-3-031-25085-9_25.

[33] G. Ying, D. Zhang, Z. Ou, X. Wang, and Z. Zheng. Temporal adaptive bidirectional bridging for RGB-D tracking. *Pattern Recognition*, 158:111053, 2025. DOI: https://doi.org/10.1016/j.patcog.2024.111053.

[34] M. Wang, F. Shi, X. Cheng, and S. Chen. Prior knowledge-driven hybrid prompter learning for RGB-event tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(9):8679–8691, 2025. DOI: https://doi.org/10.1109/TCSVT.2025.3559614.

[35] T. Zhang, Q. Zhang, K. Debattista, and J. Han. Cross-modality distillation for multi-modal tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5847–5865, 2025. DOI: https://doi.org/10.1109/TPAMI.2025.3555485.

[36] C. Sun, J. Zhang, Y. Wang, H. Ge, Q. Xia, B. Yin, and X. Yang. Exploring historical information for RGBE visual tracking with mamba. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6500–6509, 2025. DOI: https://doi.org/10.1109/CVPR52734.2025.00609.

[37] S. Chan, Z. Li, W. Li, S. Lu, C. Shen, and X. Zhang. SMSTracker: tri-path score mask sigma fusion for multi-modal tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4766–4775, 2025. DOI: https://doi.org/10.1109/ICCV51701.2025.00453.

[38] G. Zhang, Q. Liang, Z. Mo, N. Li, and B. Zhong. Visual adapt for RGBD tracking. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 9391–9395, 2024. DOI: https://doi.org/10.1109/ICASSP48485.2024.10447728.

[39] H. Wang, X. Liu, Y. Li, M. Sun, D. Yuan, and J. Liu. Temporal adaptive RGBT tracking with modality prompt. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5436–5444, 2024. DOI: https://doi.org/10.1609/aaai.v38i6.28352.

[40] Y. Chen and L. Wang. eMoE-tracker: environmental MoE-based transformer for robust event-guided object tracking. *IEEE Robotics and Automation Letters*, 10(2):1393–1400, 2025. DOI: https://doi.org/10.1109/LRA.2024.3518305.

[41] Z. Wu, J. Zheng, X. Ren, F.-A. Vasluianu, C. Ma, D.P. Paudel, L. Van Gool, and R. Timofte. Single-model and any-modality for video object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19156–19166, 2024. DOI: https://doi.org/10.1109/CVPR52733.2024.01812.

[42] Y. Ji, Q. Zhao, Y. Liang, and J. Wu. SNNPTrack: spiking neural network based prompt for high-accuracy RGBE tracking. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2025. DOI: https://doi.org/10.1109/ICASSP49660.2025.10889195.

[43] Z. Tang, T. Xu, X.-J. Wu, and J. Kittler. M3Track: meta-prompt for multi-modal tracking. *IEEE Signal Processing Letters*, 32:1705–1709, 2025. DOI: https://doi.org/10.1109/LSP.2025.3558116.

Author biographies



Xingyu Tan

Xingyu Tan received her bachelor's degree from the Communication University of China, Beijing, China, in 2020. She is currently pursuing a master's degree with the School of Intelligent Engineering and Automation, Beijing University of Posts and Telecommunications, Beijing, China. Her research interests include computer vision and machine learning, particularly multimodal object tracking.



Chun Liu

Chun Liu received the B.S. degree in mechanical design, manufacturing and automation and the M.S. degree in mechatronic engineering from China University of Geosciences, Beijing, China, in 2006 and 2009, respectively, and the Ph.D. degree in measurement and control technology from the University of Kassel, Kassel, Germany, in 2014. She is a Lecturer with the School of Intelligent Engineering and Automation, Beijing University of Posts and Telecommunications. Her research interests include deep learning, intelligent computation, computer vision, and data mining.



Qing Song

Qing Song received the Ph.D. degree from Tianjin University, Tianjin, China, in 2006. She is a Scientific Researcher with Beijing University of Posts and Telecommunications, where she is engaged in computer vision technology. She founded the Pattern Recognition and Intelligent Vision Laboratory (PRIV). She is in charge of national, provincial, and ministerial projects and enterprise cooperation projects, and has published more than 100 academic papers in international journals and conferences.



Mengjie Hu

Mengjie Hu received the Ph.D. degree from Beihang University, Beijing, China, in 2017. She is a Lecturer with the School of Intelligent Engineering and Automation, Beijing University of Posts and Telecommunications, Beijing, China. Her research interests are primarily in computer vision and machine learning, especially visual object tracking and 3D scene understanding.