LONG-TERM VISUAL OBJECT TRACKING WITH EVENT CAMERAS: AN ASSOCIATIVE MEMORY AUGMENTED TRACKER AND A BENCHMARK DATASET

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing event stream based trackers undergo evaluation on short-term tracking datasets, however, the tracking of real-world scenarios involves long-term tracking, and the performance of existing tracking algorithms in these scenarios remains unclear. In this paper, we first propose a new long-term, large-scale frame-event visual object tracking dataset, termed FELT. It contains 1,044 long-term videos that involve 1.9 million RGB frames and event stream pairs, 60 different target objects, and 14 challenging attributes. To build a solid benchmark, we retrain and evaluate 23 baseline trackers on our dataset for future work to compare. In addition, we propose a novel Associative Memory Transformer based RGB-Event long-term visual tracker, termed AMTTrack. It follows a one-stream tracking framework and aggregates the multi-scale RGB/event template and search tokens effectively via the Hopfield retrieval layer. The framework also embodies another aspect of associative memory by maintaining dynamic template representations through an associative memory update scheme, which addresses the appearance variation in long-term tracking. Extensive experiments on FELT, FE108, VisEvent, and COESOT datasets fully validated the effectiveness of our proposed tracker. Both the dataset and source code will be released upon acceptance.

1 INTRODUCTION

Visual object tracking aims to locate the specified object using an adaptively adjusted bounding box. Usually, the visual trackers are developed and deployed based on RGB cameras, which record the scenario using video frames. As we know, RGB cameras are easily influenced by challenging factors, such as low illumination, fast motion, and heavy occlusion. Thus, RGB cameras based tracking is still a very challenging task even nowadays.

To address these issues, researchers exploit new sensors to replace or assist RGB cameras for more robust tracking. Among them, bio-inspired event cameras (e.g., DVS346, CeleX-V, Prophesee) have drawn increasing attention in recent years for their unique imaging principles. Different from RGB cameras, which record all pixel intensities in the scene synchronously and output video frames at a fixed frame rate (e.g., 30 frames per second), event cameras emit a spike for each pixel only when the light intensity variation exceeds a threshold. Event pixel values (-1 or 1) are independently and asynchronously recorded, with -1 or 1 (corresponding to an OFF or ON event) denoting a decrease or increase in light intensity. As noted in Gallego et al. (2020), event cameras outperform RGB cameras in *high dynamic range*, *low energy consumption*, and *high temporal resolution*. Thus, they can be used in low-illumination and fast-motion tracking scenarios. Even with these advantages, relying solely on event cameras performs poorly in some scenes, such as static objects, as only motion signals can be sensed, and stationary objects can't be captured. Event cameras also miss color and fine-grained texture information. Fortunately, this is where RGB cameras excel. Therefore, jointly leveraging RGB and event cameras has great potential for robust visual tracking.

Based on this idea, there are already several datasets proposed for event-based tracking, including FE108 (Zhang et al., 2021), VisEvent (Wang et al., 2023), COESOT (Tang et al., 2022), EventVOT (Wang et al., 2024), and CRSOT (Zhu et al., 2024). Specifically, FE108 is collected using a gray-scale DVS event camera in the indoor scenarios, and EventVOT is a pure event-based track-

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

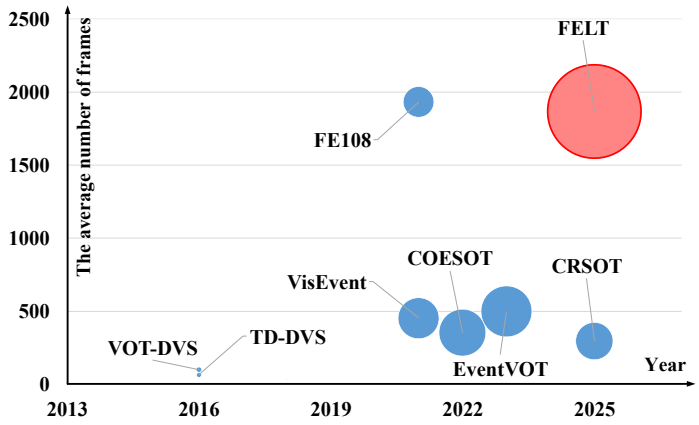


Figure 1: comparison of existing event-based datasets. The size of the bubbles represents the total number of frames.

ing dataset. VisEvent, COESOT, and CRSOT are three large-scale tracking datasets; however, their samples are short-term. Conducting long-term frame-event single object tracking has the following advantages and benefits: *1). Long-term RGB-Event tracking is closer to real-world application scenarios:* real surveillance and security systems typically require continuous tracking over extended periods, and military reconnaissance and wildlife monitoring applications need long-term tracking capabilities. *2). Greater technical challenges:* it can test the robustness of multi-modal tracking algorithms under long-term occlusion, verify the adaptivity to target appearance changes (scale, pose, illumination), and evaluate the handling capability of target disappearance and reappearance. In addition, the release of long-term multimodal tracking datasets can also *3). advance research on efficient tracking algorithms*, move beyond frame-by-frame fusion limitations, and reduce the power consumption of multimodal tracking algorithms.

Based on the aforementioned observations and reflections, in this work, we propose the first **F**rame-**E**vent **L**ong-**T**erm single object tracking benchmark dataset, termed **FELT**. It contains 1,044 videos and 1,949,680 RGB frames and event stream pairs and has become the largest frame-event tracking dataset to date, as shown in Table 1 and Fig. 1. 14 challenging attributes (e.g., illumination variation, fast motion, out of view, and occlusion) and 60 classes of target objects are involved in this dataset. In addition, we retrain and evaluate 23 baseline trackers on our dataset to build a high-quality benchmark. We believe the dataset and benchmark results build a good platform for the development of long-term frame-event single object tracking.

In this paper, we also propose a novel associative memory augmented Transformer for the long-term RGB-Event visual object tracking, termed AMTTrack. The key insight lies in leveraging associative memory to explore the correlations between features at different levels and modalities, thereby optimizing more expressive feature representations. Meanwhile, through associative memory, it aggregates diverse appearance representations of target objects during the tracking process, further alleviating the challenge of modeling object appearance in long-term tracking. More specifically, we input the template and search token sequence corresponding to the RGB and Event frames after the projection layer into the associative memory augmented Transformer backbone network for simultaneous multi-level feature extraction and interactive learning. Then, the tracking head takes the fused search tokens for target localization. In addition, we store the target object obtained in the tracking procedure and retrieve dynamic RGB/Event templates to adapt to appearance variation via a Hopfield layer. An overview of our proposed AMTTrack can be found in Fig. 2.

To sum up, the main contributions of this work can be summarized from the following three perspectives:

- We propose a large-scale and long-term RGB-Event visual object tracking dataset, termed FELT, which contains 1,044 video sequences and benchmark 23 baselines for future works to compare.
- We propose a novel associative memory augmented RGB-Event visual object tracking framework, termed AMTTrack. The expressive capability of the Transformer backbone

is further enhanced through a carefully designed hierarchical multi-modal feature interaction mechanism.

- Extensive experiments on four large-scale benchmark tracking datasets (FELT, FE108, VisEvent, COESOT) fully verify the effectiveness of our proposed tracker.

2 Related Works

2.1 Event Camera based Tracking

As a newly arising research topic in computer vision, Event-based tracking (Zhang et al., 2021) quickly draws more and more attention. Huang et al. (2018) propose RGB-CeleX sensor fusion for tracking, mining candidate regions. Cross-Attention Transformer (CMT) (Wang et al., 2023) is proposed to enable feature interaction between RGB and event data for tracking. AFNet (Zhang et al., 2023) introduces a multi-modal alignment and fusion module, which adaptively merges complementary information from event streams and RGB frames while attenuating noise. HDETrack (Wang et al., 2024) utilizes hierarchical knowledge distillation to transfer multi-modal/view knowledge to the single-modal event stream, enabling high-speed and low-latency visual tracking. CEUTrack (Tang et al., 2022) employs a unified Transformer backbone to model the relationship between RGB frames and event voxels, facilitating effective feature interaction. Zhu et al. (2023b) introduce a mask modeling strategy for cross-modal interaction between RGB frames and event streams. DANet (Fu et al., 2023) craft an event-based interference sensing tracker through the integration of Transformer modules and Siamese architecture. ViPT (Zhu et al., 2023a) incorporates learnable modal-relevant prompts, enhancing the adaptability of the models to diverse multi-modal tracking tasks. Zhu et al. (2022) directly process event clouds and predict the motion-aware target likelihood. STNet (Zhang et al., 2022) introduces a cross-domain attention fusion algorithm. SDSTrack (Hou et al., 2024) uses complementary masked patch distillation to boost extreme conditions tracking. MamTrack (Sun et al., 2025) leverages Mamba for long-range historical modeling. LASTracker (Wang et al., 2026) switches between ANN and SNN branches to balance accuracy and efficiency. Different from these works, we propose a novel long-term RGB-Event tracker using the associative memory Transformer.

2.2 Modern Hopfield Networks

Hopfield network (Hopfield, 1982) proposed by John Hopfield is well known for its associative memories, and it is treated as one of the earliest artificial neural models. This neural network can associate an input with its most similar pattern, i.e., retrieve a stored memory based on partial or incomplete input. Different from the classical binary Hopfield networks, recent works (Krotov & Hopfield, 2016; Demircigil et al., 2017) demonstrate that by introducing polynomial or exponential terms into the energy function, the storage capacity can be increased significantly. Modern Hopfield networks (Ramsauer et al., 2020) achieve continuous associative memory. After that, many tasks are boosted with the help of modern Hopfield networks, such as remote sensing image generation (Xu et al., 2023), reaction template prediction (Seidl et al., 2022), immune repertoire classification (Widrich et al., 2020), nuclear fusion (Ma et al., 2024), and X-ray report generation (Wang et al., 2025b). Inspired by the strong associative memory of modern Hopfield networks, we introduce associative memory to model correlations in long-term RGB-Event tracking.

Table 1: comparison of event camera-based visual tracking datasets.

Datasets	Year	Videos	Frames	Class	Attr.	Resolution	Aim	Absent	Frame	Reality	LT	Public
VOT-DVS	2016	60	-	-	-	240 × 180	Eval	✗	✗	✗	✗	✓
TD-DVS	2016	77	-	-	-	240 × 180	Eval	✗	✗	✗	✗	✓
Ulster	2016	1	9,000	-	-	240 × 180	Eval	✗	✗	✓	✓	✗
EED	2018	7	234	-	-	240 × 180	Eval	✗	✗	✓	✗	✓
FE108	2021	108	208,672	21	4	346 × 260	Train & Eval	✗	✗	✓	✓	✓
VisEvent	2021	820	371,127	-	**17**	346 × 260	Train & Eval	✓	✓	✓	✗	✓
COESOT	2022	**1,354**	478,721	**90**	**17**	346 × 260	Train & Eval	✓	✓	✓	✗	✓
EventVOT	2023	1,141	569,359	19	14	1280 × 720	Train & Eval	✓	✗	✓	✗	✓
CRSOT	2025	1,030	304,974	-	**17**	**1280 × 800**	Train & Eval	✓	✓	✓	✗	✓
FELT (Ours)	2026	1,044	**1,949,680**	60	14	346 × 260	Train & Eval	✓	✓	✓	✓	✓

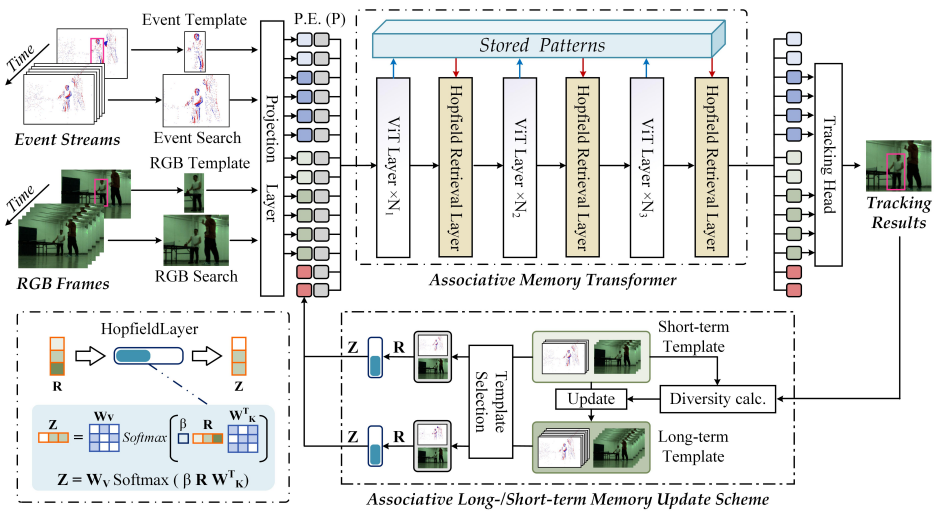


Figure 2: an overview of our proposed associative memory augmented RGB-Event visual tracking framework. Our tracker (termed AMTTrack) is composed of three parts: Input Embedding Layer, Associative Memory Transformer based on Retrieval Augmentation, and Associative Long-/Short-term Memory Update Scheme.

3 FELT Single Object Tracking Dataset

3.1 Dataset Design and Collection

The FELT SOT dataset collection follows these principles: *1). Long-term*: Each video has at least 1000 frames and event streams. *2). Large-scale*: To support data-hungry DNNs, we aim for the largest frame-event dataset. *3). Various Challenges*: Videos should reflect key frame-event tracking challenges. *4). Dual-Modality*: The dataset includes RGB and Event data for event-only and fusion tasks. *5). Incomplete Information*: Accounts for modality-specific incompleteness, e.g., sparse event and weakened RGB data.

The DVS346 event camera, which outputs the RGB frames (346×260) and event streams simultaneously, is adopted for FELT dataset collection. The two modalities are aligned well from a hardware perspective on both spatial and temporal views. When recording our videos, we follow the design principles and challenging attributes. For the bounding box annotation, we employ a professional data company to annotate and check for several rounds to ensure quality.

3.2 Statistical Analysis

When collecting the FELT SOT dataset, we consider 14 attributes (detailed in the supplementary material) for the long-term frame-event tracking task. Specifically, the ST (Small Target) and illumination related attributes (OE (Over Exposure), IV (Illumination Variation), LI (Low Illumination)) are all considered. We also define the motion-based attributes, like the NMO (No MOtion) and FM (Fast Motion). The OV (Out of View) is considered for the long-term tracking task. We believe these defined challenging attributes will be a better platform for the evaluation of future trackers. As shown in Table 1, our dataset contains 1,044 videos with 1,949,680 RGB frames, covering 60 target object categories in indoor and outdoor scenarios. We split these videos into training and testing subsets, which contain 730 and 314 videos, respectively.

4 Methodology

4.1 Overview

As shown in Fig. 2, the input RGB and Event templates and search areas are embedded into token sequences via projection encoding and position embedding and then fed into the unified ViT (Doso-

vitskiy et al., 2020) backbone for relational modeling. We embed Hopfield retrieval layers to enhance multi-level feature extraction and representation. Finally, the tracking head uses fused search features for object positioning. Tracking results can be stored in template memory, and templates are dynamically updated during tracking.

4.2 Associative Memory Augmented Hopfield

This subsection elaborates on the implementation of the Associative Memory Augmented Hopfield module in our RGB-Event visual tracking framework. The module employs the Hopfield network for feature retrieval, thereby enhancing the model's hierarchical feature representation.

Modern Hopfield Networks. According to Ramsauer et al. (2020), the modern Hopfield networks converge to a steady-state by minimizing an energy function corresponding to stored memory patterns. As shown in Fig. 3 (b), modern Hopfield networks exhibit a connection with attention mechanisms, which can be formally expressed as follows:

$$\boldsymbol{Z} = \boldsymbol{Y}\boldsymbol{W}_V \cdot \mathrm{softmax}\left(\beta \boldsymbol{R}\boldsymbol{W}_Q \boldsymbol{W}_K^T \boldsymbol{Y}^T\right), \tag{1}$$

where $\boldsymbol{Z}$ denotes the result patterns, $\boldsymbol{R}$ represents the state patterns, and $\boldsymbol{Y}$ means the stored patterns. The projection matrices $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$, and $\boldsymbol{W}_V$ transform these patterns. β controls the retrieval sharpness: higher β retrieves individual patterns, while lower β may blend multiple patterns.

Hopfield Embedding in Transformer Layers. Our network consists of a 12-layer Transformer backbone. As shown in Fig. 3 (a), we embed Hopfield layers into specific layers of the Transformer. These Hopfield layers capture different semantic levels during feature extraction and interaction. Let $\mathcal{L} = \{l_1, l_2, \ldots, l_k\}$ represent the set of Hopfield layers in a Transformer architecture, where each l_i denotes the positional index of the i-th Hopfield layer within the 12-layer Transformer. For each $l_i \in \mathcal{L}$, we denote features from preceding layers $\mathcal{P}_i = \{p_{i1}, p_{i2}, \cdots, p_{im}\}$ as Hopfield's memory bank $\boldsymbol{Y}_i$, with p_{ij} as the j-th preceding layer's index. The output of the current transformer layer denoted as $\boldsymbol{R}_i$ will be utilized for retrieval of the stored patterns $\boldsymbol{Y}_i$.

Cross-Modal Retrieval. With dual-modal token sequences, we introduce a cross-modal retrieval mechanism, as shown in Fig. 3 (a), which fully explores the complementary information between different modalities. In our cross-modal retrieval process, we leverage the Hopfield layer to enhance the feature representation. Specifically, for each Hopfield layer $l_i \in \mathcal{L}$, when $\boldsymbol{R}_v^i$ is the RGB search feature of the l_i-th layer, we employ Event features from the layers in $\mathcal{P}_i$ as $\boldsymbol{Y}_e^i$ for retrieval. Conversely, for $\boldsymbol{R}_e^i$ (Event search feature), the retrieval process is similar. Mathematically, the cross-modal retrieval process can be formulated as follows:

$$\begin{cases} \boldsymbol{Z}_v^i = Hopfield(\boldsymbol{R}_v^i, \boldsymbol{Y}_e^i) + \boldsymbol{R}_v^i, \\ \boldsymbol{Z}_e^i = Hopfield(\boldsymbol{R}_e^i, \boldsymbol{Y}_v^i) + \boldsymbol{R}_e^i, \end{cases} \tag{2}$$

where $\boldsymbol{Z}_v^i$ and $\boldsymbol{Z}_e^i$ are the output features after cross-modal retrieval, and $Hopfield(\cdot, \cdot)$ represents the Hopfield layer.

Specifically, through the cross-modal retrieval mechanism described above, it enables our model to effectively integrate multi-modal information, enhancing tracking accuracy and robustness, especially in complex scenarios with insufficient single-modality data. Moreover, the retrieved patterns from multi-layer dual-modal features further integrate local details and global context from different Transformer layers.

4.3 Associative Template Update

Long-term visual tracking may cause drastic object appearance changes, making it hard to localize the target accurately with only the initial static template. We introduce an associative memory-based dynamic template updating strategy, which maintains a short-term template memory and a long-term template memory, and design a dynamic template updating mechanism based on appearance diversity matching for dynamically updating the template memory during short- and long-term visual tracking.

Long-/Short-term Template Memory. Inspired by THOR (Sauer et al., 2019) and MambaEVT (Wang et al., 2025a), we use short-term (ST) and long-term (LT) memories to store templates

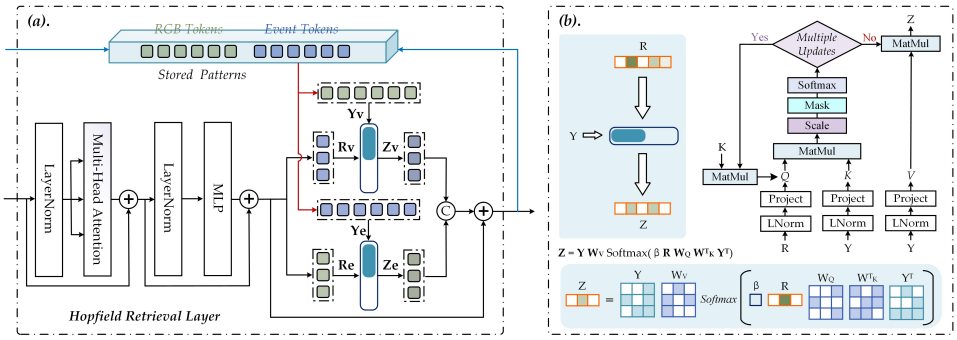


Figure 3: detailed architecture of (a). Hopfield Retrieval Layer; (b). Modern Hopfield Network.

Algorithm 1 algorithm for dynamic template update

Input: $\mathbf{z}_s$, $\mathbf{z}_d$, **score**
Parameter: I_{su}, I_{lu}, θ, τ
Output: Dynamic templates $\mathbf{z}$

Init: Use $\mathbf{z}_s$ to init short/long-term mem.; $n \leftarrow 0$
for each frame **do**
  $n \leftarrow n + 1$
  if score $> \tau$ and $n \bmod I_{su} = 0$ **then**
    div_scale $\leftarrow$ Update ST mem.($\mathbf{z}_d$)
    if $n \bmod I_{lu} = 0$ **then**
      Update LT mem.($\mathbf{z}_d$, div_scale, θ)
    end if
  end if
  $\mathbf{z}_d^{LT}, \mathbf{z}_d^{ST} \leftarrow$ ResampleTemplate{ST, LT}
  $\mathbf{z}_d^{LT}, \mathbf{z}_d^{ST} \leftarrow$ HopfieldLayer($\mathbf{z}_d^{LT}, \mathbf{z}_d^{ST}$)
  $\mathbf{z} \leftarrow$ Concat($[\mathbf{z}_d^{LT}, \mathbf{z}_d^{ST}]$)
end for

as cropped images from previous predictions. The ST maintains a dynamic template bank, updated when two conditions are met: (1) the predefined short-term update interval arrives, and (2) the response map's maximum confidence score surpasses a specified threshold. Based on these rules, we introduce an appearance-diversity-based long-term template update mechanism. It enriches the LT with diverse features to adapt to object appearance variability.

Associative Hopfield Augmented Template. In complex scenarios, even optimal template updates can add low-quality templates to memory, ultimately leading to tracking performance degradation. To address this, we introduce association memory-augmented dynamic templates (ATU) to optimize their representation. As shown in Fig. 2, the Hopfield layer maintains learnable prototypes as static stored patterns through its bias weights. Dynamic template features are used as state patterns $\boldsymbol{R}$ for retrieval; therefore, no additional input-dependent stored patterns $\boldsymbol{Y}$ are required. The Hopfield layer can be expressed as $\boldsymbol{Z} = \boldsymbol{W}_V \cdot \mathrm{softmax}(\beta \boldsymbol{R}\boldsymbol{W}_K^T)$. Specifically, during training, dynamic templates are used to optimize the learnable prototypes. During inference, templates sampled from memory serve as state patterns $\boldsymbol{R}$. The Hopfield layer outputs refined templates $\boldsymbol{Z}$, enhancing their representation. In ATU, the Hopfield layer acts as a denoising reconstructor where dynamic features from ST/LT memories serve as state patterns $\boldsymbol{R}$. It reconstructs denoised features via associative retrieval against learnable prototypes, enhancing template quality and tracking robustness.

Update Process. The dynamic template update process is detailed in Algorithm 1. Inputs: initial static template $\mathbf{z}_s$, dynamic template $\mathbf{z}_d$, response **score**. Parameters: short-term update interval I_{su}, long-term update interval I_{lu}, similarity lower bound θ, response threshold τ.

4.4 Tracking Head and Loss Function

We input the dual-modal fused search features into the tracking head, which has the same configuration as OSTrack (Ye et al., 2022), for object localization and predicting the bounding box. During training, the loss function includes L1 loss, GIOU loss (Rezatofighi et al., 2019), and weighted focal

Table 2: experimental results (SR/PR) on FE108 dataset.

SiamRPN	SiamBAN	SiamFC++	KYS	CLNet	CMT-MDNet	ATOM
21.8/33.5	22.5/37.4	23.8/39.1	26.6/41.0	34.4/55.5	35.1/57.8	46.5/71.3
DiMP	**PrDiMP**	**CMT-ATOM**	**CEUTrack**	**ViPT**	**MamTrack**	**Ours**
52.6/79.1	53.0/80.5	54.3/79.4	55.58/84.46	65.8/93.8	**66.4**/94.2	65.6/**95.9**

Table 3: experimental results on FELT dataset.

Trackers	Source	SR	PR	NPR	FPS	Trackers	Source	SR	PR	NPR	FPS
01. Stark	ECCV22	52.7	67.9	62.8	42	**13. EVPTrack**	AAAI24	53.8	68.7	64.8	45
02. OSTrack	ECCV22	52.3	65.9	63.3	75	**14. AQATrack**	CVPR24	54.0	**69.1**	64.7	41
03. MixFormer	CVPR22	53.0	67.5	63.8	36	**15. SDSTrack**	CVPR24	53.7	66.4	64.1	28
04. AiATrack	ECCV22	52.2	66.7	62.8	31	**16. UnTrack**	CVPR24	53.6	66.0	63.9	12
05. SimTrack	ECCV22	49.7	63.6	59.8	82	**17. FERMT**	ECCV24	51.8	66.1	62.9	70
06. GRM	CVPR23	52.1	65.6	62.9	38	**18. LMTrack**	AAAI25	50.9	63.9	61.8	72
07. ROMTrack	ICCV23	51.8	65.8	62.7	64	**19. AsymTrack**	AAAI25	51.9	66.7	62.0	104
08. ViPT	CVPR23	52.8	65.3	63.1	29	**20. UNTrack**	CVPR25	50.0	63.9	61.6	23
09. SeqTrack	CVPR23	52.7	66.9	63.4	31	**21. SUTrack**	AAAI25	**55.1**	68.9	65.2	44
10. ARTrackv2	CVPR24	52.3	65.2	62.8	42	**22. XTrack**	ICCV25	**55.1**	67.6	65.3	10
11. HIPTrack	CVPR24	51.6	65.6	62.2	39	**23. SpikeTrack**	CVPR26	52.5	67.9	63.5	30
12. ODTrack	AAAI24	52.2	66.0	63.5	57	**24. Ours**	-	54.8	67.9	**65.7**	31

loss (Law & Deng, 2018), with $\lambda_{\mathcal{L}_1} = 5.0$ and $\lambda_{GIoU} = 2.0$:

$$\mathcal{L} = \lambda_{\mathcal{L}_1}\mathcal{L}_1 + \lambda_{GIoU}\mathcal{L}_{GIoU} + \mathcal{L}_{focal}, \tag{3}$$

5 Experiments

5.1 Datasets and Evaluation Metric

To evaluate our proposed framework, extensive experiments are conducted on four benchmark datasets: **FE108**, **FELT**, **COESOT**, and **VisEvent**. The above datasets contain different RGB-Event challenge scenarios and cover various complex factors that arise in the short-term and long-term tracking process. For evaluation, we employ three widely used metrics: **Success Rate (SR)**, **Precision (PR)**, and **Normalized Precision (NPR)**.

5.2 Comparison on Public Benchmarks

• **Results on FE108 Dataset.** FE108 (Zhang et al., 2021) averages 1,932 frames per video, making it suitable for evaluating long-term tracking scenarios. As shown in Table 2, we report the results of our trackers and other SOTA methods on FE108. Our tracker achieves competitive performance, with the precision rate surpassing all comparable trackers.

• **Results on FELT Dataset.** FELT dataset contains 1044 long-term video sequences with an average of about 1867 frames per sequence. We retrain multiple SOTA trackers and report their performance on the FELT dataset. As shown in Table 3, compared to our baseline OSTrack, our tracker with the associative memory-based Hopfield layer and template update improves the performance on the FELT dataset, achieving SR 54.8, PR 67.9, and NPR 65.7. Notably, our method achieves the highest NPR. However, the PR metric is lower than that of AQATrack and EVPTrack, which leverage temporal information to capture historical features.

• **Results on COESOT Dataset.** COESOT (Tang et al., 2022) is a large-scale benchmark dataset for color-event tracking, including 1,354 video sequences. As shown in Table 4, our method outperforms other SOTA trackers on multiple metrics, with SR 68.8 and PR 82.9.

• **Results on VisEvent Dataset.** VisEvent (Wang et al., 2023) contains 820 video sequences and has 17 challenging factors. As shown in Table 4, our method is trained and evaluated exclusively on VisEvent, achieving an SR of 60.1 and the best PR of 78.1. Its SR is only 0.2 points lower than SEATrack, while its PR surpasses SEATrack by 1.0 point.

Table 4: experimental results on COESOT and VisEvent.

Trackers	Source	COESOT		VisEvent	
		SR	PR	SR	PR
01. Stark	ICCV21	56.0	67.7	44.6	61.2
02. TransT	ECCV22	60.5	72.4	47.4	65.0
03. OSTrack	ECCV22	59.0	70.7	53.4	69.5
04. ViPT	CVPR23	68.3	81.0	59.2	75.8
05. SDSTrack	CVPR24	66.7	79.7	59.7	76.7
06. UnTrack	CVPR24	67.9	80.9	59.7	76.3
07. CMDTrack	TPAMI25	65.7	74.8	-	-
08. LMTrack	AAAI25	58.4	71.1	-	-
09. SEATrack	CVPR26	-	-	**60.3**	77.1
10. LASTracker	PR26	62.6	72.3	52.9	70.4
11. Ours	-	**68.8**	**82.9**	60.1	**78.1**

Table 5: ablation studies of components and hyperparameters on COESOT.

#	# Settings	SR	PR	NPR	#	# Settings	SR	PR	NPR
Component	Baseline	63.2	77.1	76.1	**AMAH-β**	0.1	63.8	78.2	77.1
	+ATU	64.1	78.7	77.5		**0.25**	**63.9**	**78.4**	**77.1**
	+AMAH	63.9	78.4	77.1		1.0	63.5	77.6	76.6
	+ATU+AMAH	**64.8**	**79.4**	**78.2**		4.0	62.6	76.7	75.9
ATU-β	1.0	64.2	78.5	77.4	**Layers**	$[3, 7, 11]$	63.4	77.5	76.6
	4.0	**64.8**	**79.4**	**78.2**		$[5, 8, 11]^1$	63.6	77.9	76.7
	16.0	64.5	78.9	78.0		$\mathbf{[5, 8, 11]^2}$	**63.9**	**78.4**	**77.1**
ST/LT	3 / 6	64.6	79.1	77.9	**RT**	ST & ST	64.6	79.2	78.0
	5 / 10	**64.8**	**79.4**	**78.2**		LT & LT	64.7	79.3	78.1
	7 / 14	64.5	79.1	77.9		**ST & LT**	**64.8**	**79.4**	**78.2**
ATU-Hopfield	✗	63.9	78.0	77.3	**Model**	Baseline	63.2	77.1	76.1
	✓	**64.8**	**79.4**	**78.2**		MLP	63.2	77.3	76.2
RI/UI	5 / 10	64.7	79.3	78.2		Hopfield Pooling	63.2	77.5	76.7
	5 / 20	64.6	79.1	77.9		Hopfield	63.5	77.4	76.8
	10 / 20	**64.8**	**79.4**	**78.2**		**Cross Hopfield**	**63.9**	**78.4**	**77.1**
	20 / 40	64.5	79.1	77.8					

5.3 Ablation Study

• **Analysis on Different Weights.** As shown in Table 6, we evaluate our method using two pretrained weights (MAE and OSTrack). Compared to MAE pretrained weights, OSTrack pretrained weights lead to remarkable improvements. Across the FE108, FELT, COESOT, and VisEvent datasets, SR, PR, and NPR all increase. This consistent improvement highlights the importance of strong RGB pretrained weights. Fine-tuned on the pretrained OSTrack, our model acquires rich visual representations and tracking behaviors, facilitating effective knowledge transfer and enabling fast adaptation across different datasets.

• **Effectiveness of Proposed Components.**

We conduct ablation experiments on each module using MAE pretrained weights. Table 5 shows the effectiveness of Associative Template Update (ATU) and Associative Memory Augmented Hopfield (AMAH). ATU improves performance on both datasets by adapting templates to target changes, while AMAH also enhances performance by boosting dual-modal fusion in the backbone.

• **Analysis on Associative Template Update.**

As shown in Table 5, we analyze parameters of associative memory-based dynamic template updates. First, we compare using and not using the Hopfield layer to enhance dynamic templates. The

Table 6: performance of different weights for each dataset.

Datasets	MAE			OSTrack		
	SR	PR	NPR	SR	PR	NPR
FE108	63.0	94.4	68.1	65.6	95.9	71.8
FELT	53.5	66.9	64.6	54.8	67.9	65.7
COESOT	64.8	79.4	78.2	68.8	82.9	81.8
VisEvent	56.4	74.6	70.0	60.1	78.1	73.9

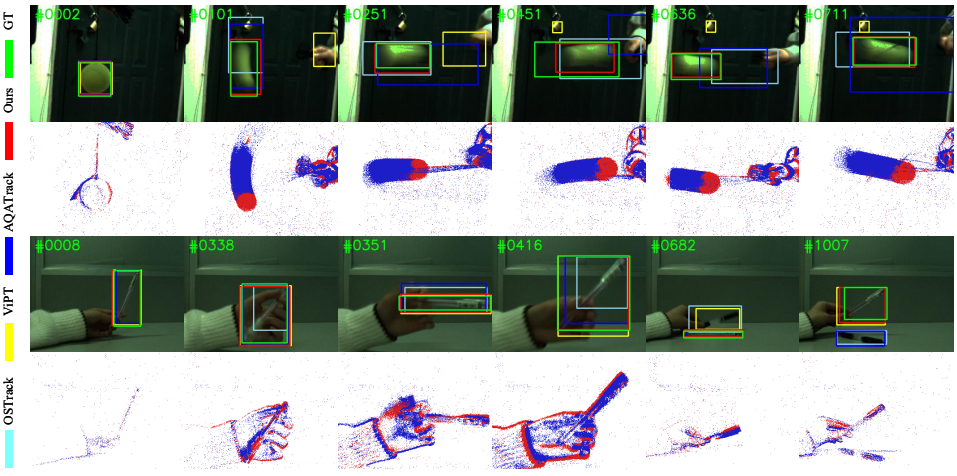


Figure 4: tracking results of AMTTrack and other trackers.

results show significant improvement with the Hopfield layer. Then, we analyze multiple template update parameters: β in the Hopfield layer, short- and long-term template memory lengths (ST and LT), template resampling and update intervals (RI and UI), and the resampling choice between short- and long-term template memory (RT).

• **Analysis on Associative Memory Augmented Hopfield.** Table 5 shows the analysis of Associative Memory Augmented Hopfield. To demonstrate the effectiveness of Hopfield layer's cross-modal retrieval, we compare MLP, Hopfield Pooling, and the Hopfield Layer without cross-modal retrieval, keeping the selected layer consistent across experiments. The results show our Hopfield layer's cross-modal retrieval performs better. As shown in Table 5, we analyze and compare cross-modal Hopfield retrieval layer hyperparameters. Setting β to 0.25, which superposes multiple stored patterns as the result, yields the best performance and enables multi-layer contextual feature integration. The selection of stored patterns and preceding layer features in the Hopfield layer is crucial. We compare three schemes and select $[5, 8, 11]^2$ ($\mathcal{L} = \{5, 8, 11\}$, $\mathcal{P}_5 = \{1, 3\}$, $\mathcal{P}_8 = \{4, 6\}$, $\mathcal{P}_{11} = \{7, 9\}$) for our Hopfield retrieval layer. Details are provided in the supplementary material.

• **Efficiency and Parameter Analysis.** The full model increases parameters from 92.52M to 101.16M and FLOPs from 89.55G to 99.16G, incurring only 9.3% and 10.7% overheads, respectively. Inference runs at 31 FPS on FELT.

5.4 Visualization and Limitation Analysis

Beyond the quantitative analysis, we also provide a visualization to help readers better understand our tracking algorithm. Fig. 4 demonstrates the robust performance of our tracker against other SOTA visual trackers, confirming its effectiveness in RGB-Event long-term visual tracking.

While our method performs well on multiple datasets, some metrics can be optimized. On FELT, our PR is outperformed by temporal model-based trackers. Future work will leverage frame-to-frame temporal cues or explore dense temporal resolution within event data to enhance long-term tracking.

6 Conclusion

In this work, we propose a new long-term and large-scale frame-event single object tracking dataset, termed FELT. It contains 1,044 videos and 1,949,680 RGB frames and event stream pairs and is also the largest frame-event tracking dataset to date. To build a comprehensive benchmark, we also retrain and evaluate multiple baseline trackers on our newly proposed FELT SOT dataset, which will be a better platform for future works to compare. In addition, we propose AMTTrack, a novel associative memory-augmented transformer-based tracker. It utilizes modern Hopfield layers to jointly optimize multi-modal feature extraction and representation within the backbone network while refining dynamic templates, thereby enhancing the quality of multi-modal long-term tracking.

AI USE STATEMENT

In this work, we used generative AI tools (large language models) to aid and polish the writing of this manuscript. We have not used generative AI tools for any other tasks, including research idea generation, experiments, code, figures, and data analysis. We have reviewed all AI-assisted text and take responsibility for the final content of this work.

REFERENCES

Mete Demircigil, Judith Heusel, Matthias Löwe, Sven Upgang, and Franck Vermet. On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, 168(2):288–299, 2017.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Yingkai Fu, Meng Li, Wenxi Liu, Yuanchen Wang, Jiqing Zhang, Baocai Yin, Xiaopeng Wei, and Xin Yang. Distractor-aware event-based tracking. *IEEE Transactions on Image Processing*, 32: 6129–6141, 2023.

Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020.

John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26551–26561, 2024.

Jing Huang, Shizheng Wang, Menghan Guo, and Shoushun Chen. Event-guided structured output tracking of fast-moving objects using a celex sensor. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(9):2413–2417, 2018.

Dmitry Krotov and John J Hopfield. Dense associative memory for pattern recognition. *Advances in neural information processing systems*, 29, 2016.

Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 734–750, 2018.

Qingchuan Ma, Shiao Wang, Tong Zheng, Xiaodong Dai, Yifeng Wang, Qingquan Yang, and Xiao Wang. Exploiting memory-aware q-distribution prediction for nuclear fusion via modern hopfield network. In *International Conference on Brain Inspired Cognitive Systems*, pp. 104–114. Springer, 2024.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio 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*, pp. 658–666, 2019.

Axel Sauer, Elie Aljalbout, and Sami Haddadin. Tracking holistic object representations. *arXiv preprint arXiv:1907.12920*, 2019.

Philipp Seidl, Philipp Renz, Natalia Dyubankova, Paulo Neves, Jonas Verhoeven, Jorg K Wegner, Marwin Segler, Sepp Hochreiter, and Gunter Klambauer. Improving few-and zero-shot reaction template prediction using modern hopfield networks. *Journal of chemical information and modeling*, 62(9):2111–2120, 2022.

Chuanyu Sun, Jiqing Zhang, Yang Wang, Huilin Ge, Qianchen Xia, Baocai Yin, and Xin Yang. Exploring historical information for rgbe visual tracking with mamba. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6500–6509, 2025.

Chuanming Tang, Xiao Wang, Ju Huang, Bo Jiang, Lin Zhu, Jianlin Zhang, Yaowei Wang, and Yonghong Tian. Revisiting color-event based tracking: A unified network, dataset, and metric. *arXiv preprint arXiv:2211.11010*, 2022.

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.

Xiao Wang, Shiao Wang, Chuanming Tang, Lin Zhu, Bo Jiang, Yonghong Tian, and Jin Tang. Event stream-based visual object tracking: A high-resolution benchmark dataset and a novel baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19248–19257, 2024.

Xiao Wang, Chao Wang, Shiao Wang, Xixi Wang, Zhicheng Zhao, Lin Zhu, and Bo Jiang. Mambaevt: Event stream based visual object tracking using state space model. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025a.

Xiao Wang, Fuling Wang, Haowen Wang, Bo Jiang, Chuanfu Li, Yaowei Wang, Yonghong Tian, and Jin Tang. Activating associative disease-aware vision token memory for llm-based x-ray report generation. *arXiv preprint arXiv:2501.03458*, 2025b.

Zikai Wang, Siying Liu, Hanle Zheng, Sihan Wang, Yifan Hu, Hanxiao Fan, Yao Li, Hao Guo, and Lei Deng. Lastracker: A lightweight rgb-e tracking framework with ann-snn adaptive switching. *Pattern Recognition*, pp. 113623, 2026.

Michael Widrich, Bernhard Schäfl, Milena Pavlović, Hubert Ramsauer, Lukas Gruber, Markus Holzleitner, Johannes Brandstetter, Geir Kjetil Sandve, Victor Greiff, Sepp Hochreiter, et al. Modern hopfield networks and attention for immune repertoire classification. *Advances in neural information processing systems*, 33:18832–18845, 2020.

Yonghao Xu, Weikang Yu, Pedram Ghamisi, Michael Kopp, and Sepp Hochreiter. Txt2img-mhn: Remote sensing image generation from text using modern hopfield networks. *IEEE Transactions on Image Processing*, 32:5737–5750, 2023.

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

Jiqing Zhang, Xin Yang, Yingkai Fu, Xiaopeng Wei, Baocai Yin, and Bo Dong. Object tracking by jointly exploiting frame and event domain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13043–13052, 2021.

Jiqing Zhang, Bo Dong, Haiwei Zhang, Jianchuan Ding, Felix Heide, Baocai Yin, and Xin Yang. Spiking transformers for event-based single object tracking. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 8801–8810, 2022.

Jiqing Zhang, Yuanchen Wang, Wenxi Liu, Meng Li, Jinpeng Bai, Baocai Yin, and Xin Yang. Frame-event alignment and fusion network for high frame rate tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9781–9790, 2023.

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, 2023a.

Yabin Zhu, Xiao Wang, Chenglong Li, Bo Jiang, Lin Zhu, Zhixiang Huang, Yonghong Tian, and Jin Tang. Crsot: Cross-resolution object tracking using unaligned frame and event cameras. *arXiv preprint arXiv:2401.02826*, 2024.

Zhiyu Zhu, Junhui Hou, and Xianqiang Lyu. Learning graph-embedded key-event back-tracing for object tracking in event clouds. *Advances in Neural Information Processing Systems*, 35:7462–7476, 2022.

Zhiyu Zhu, Junhui Hou, and Dapeng Oliver Wu. Cross-modal orthogonal high-rank augmentation for rgb-event transformer-trackers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22045–22055, 2023b.

A Benchmark Datasets for Event-based Tracking

In this supplementary material, we provide detailed technical elaborations and additional experimental results to complement the main paper. Specifically, we present: (1) further details of the FELT dataset and existing benchmarks; (2) more details about Modern Hopfield Networks and ATU; (3) extended experimental details.

Event stream-based visual tracking has garnered increasing attention in recent years, highlighting the growing importance of event-based benchmark datasets. To uphold dataset quality, manual labeling of target boxes is often a requisite for event-based benchmark datasets. Existing widely used event-based single object tracking datasets include FE108 (Zhang et al., 2021), VisEvent (Wang et al., 2023), COESOT (Tang et al., 2022), EventVOT (Wang et al., 2024) and CRSOT (Zhu et al., 2024).

FE108 (Zhang et al., 2021) comprises 108 videos captured under consistent lighting conditions, totaling 208,672 frames for the entire dataset, averaging 1,932 frames per video. The dataset encompasses 21 distinct categories of target objects and introduces four challenging factors.

VisEvent (Wang et al., 2023) encompasses 820 video sequences, totaling 371,127 frames, with an average length of 453 frames per video. Six fundamental guidelines were adhered to during the dataset collection process to ensure high-quality, diverse, and large-scale datasets. The dataset incorporates 17 challenging factors, encompassing RGB mode challenges such as fast motion, low illumination, and exposure, as well as event mode challenges like the absence of target movement, background confusion, etc.

COESOT (Tang et al., 2022) is a large-scale benchmark dataset for color-event tracking. It comprises 1,354 video sequences encompassing 90 target object categories. The dataset spans a total of 478,721 frames, averaging 354 frames per video. The diverse video scenes include indoor, street, zoo, and various other environments. Noteworthy challenges within the dataset involve factors such as the rotation of target objects, complete occlusion, and movement of background objects.

EventVOT (Wang et al., 2024) stands as the pioneering high-resolution, large-scale, single-modal event dataset. Encompassing 1,141 video sequences, the dataset comprises a total of 569,359 frames, averaging 499 frames per video. Capturing a diverse range of scenes, including indoor, street, and stadium, the videos span both day and nighttime periods. The dataset prominently features 14 challenging factors, such as speed changes and variations in lighting conditions. Notably, all data undergo annotation by a professional labeling company to ensure the dataset's overall high quality.

CRSOT (Zhu et al., 2024) is the first unaligned frame-event dataset. It employs a hybrid camera system to capture high-resolution data (1440×1080 RGB, 1280×800 CeleX-V), with the NIR camera assisting low-light annotation. With 1,030 RGB-Event pairs and 304,974 frames, it captures DVS features, and defines 17 attributes for tracker performance evaluation. Key challenges include background clutter, background object motion, low illumination, and partial occlusion.

FELT is the **first long-term** frame-event tracking dataset and the **largest** frame-event tracking dataset to date. In the FELT dataset, we have defined 14 challenging factors, as described in Table 7. The number of videos corresponding to these attributes is visualized in Fig. 5 (a). We find that BI, OV, and VT are the top three major challenges, indicating that our dataset is relatively challenging. The distribution of the center points of the annotated bounding boxes is visualized in Fig. 5 (b).

Table 7: description of 14 attributes in our FELT dataset.

Attributes	Description
01. ST	The scale of the target object is limited
02. OE	The light intensity is very high
03. IV	The light intensity changes during tracking
04. NMO	The target object is stationary
05. ARC	The ratio of the bounding box aspect ratio varied
06. BI	The target object is influenced by the background
07. SV	The width and height of the target object changed
08. POC	Part of the target object is occluded
09. OV	The target object moves out of the view
10. LI	The videos are recorded in the dark environment
11. FOC	The target object is fully occluded by other objects
12. DEF	The shape of the target object changed
13. VT	The views of the target object vary during tracking
14. FM	The target object moves quickly

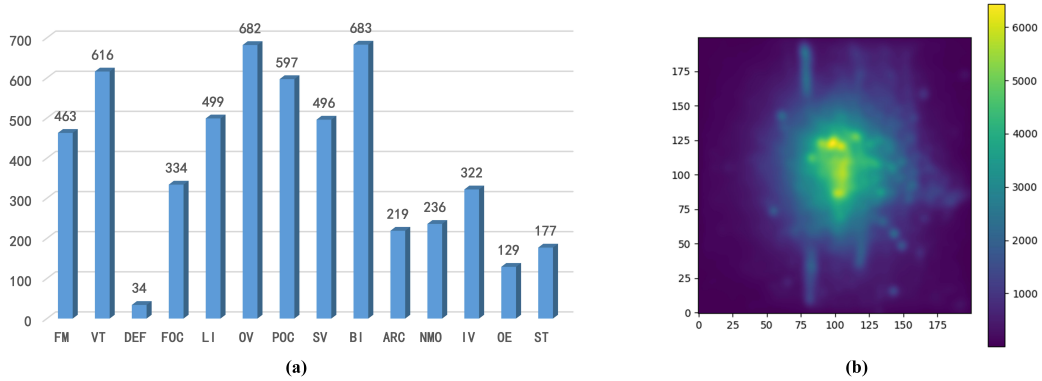


Figure 5: (a) the number of videos corresponding to specific challenging factors; (b) distribution of bounding boxes.

B METHODOLOGY

B.1 INPUT REPRESENTATION

The input of our AMTTrack framework consists of template and search area images from RGB and Event, where the sizes of the template and search area are 128×128 and 256×256, respectively. We denote the RGB frames as $\mathcal{I}_i \in \mathbb{R}^{C \times H \times W}$ and the Event frames as $\mathcal{E}_i \in \mathbb{R}^{C \times H \times W}$, where i is the index of RGB-Event image pairs. The event frames $\mathcal{E}_i$ are composed of event streams $\{e_1, e_2, \cdots, e_M\}$ stacked at a fixed interval, where $e_j = [x, y, t, p]$ represents each event point. Then, we first perform patch embedding and add position encoding on template frames and search frames to obtain template and search token sequences of RGB and Event, which are RGB template token embeddings $\mathcal{Z}_v \in \mathbb{R}^{N_z \times D}$, RGB search token embeddings $\mathcal{X}_v \in \mathbb{R}^{N_x \times D}$, Event template token embeddings $\mathcal{Z}_e \in \mathbb{R}^{N_z \times D}$, and Event search token embeddings $\mathcal{X}_e \in \mathbb{R}^{N_x \times D}$, respectively. Here, $N_z = \frac{H_z \times W_z}{P^2}$, $N_x = \frac{H_x \times W_x}{P^2}$, and $P = 16$.

B.2 MORE DETAILS ABOUT HOPFIELD

According to Ramsauer et al. (2020), the modern Hopfield networks converge to a steady state by minimizing an energy function corresponding to stored memory patterns. By adjusting the weight matrix, the network can store multiple patterns, with each pattern corresponding to a steady state. Its energy function is expressed as:

$$E(\boldsymbol{r}) = -\frac{1}{\beta} \log \sum_{i=1}^{N} \exp(\beta \boldsymbol{y}_i^\top \boldsymbol{r}) + \frac{1}{2} \|\boldsymbol{r}\|^2 + C, \tag{4}$$

where β denotes a temperature parameter, $\boldsymbol{y}_i$ represents the i-th stored pattern, $\boldsymbol{r}$ denotes the input state pattern, and N stands for the number of stored patterns. C represents a constant quadratic term. Inspired by Ramsauer et al. (2020), modern Hopfield networks exhibit a strong connection with attention mechanisms, and the update rule of the new energy function is the self-attention of transformer networks. It can be formally expressed as:

$$\boldsymbol{r}^{(t+1)} = \boldsymbol{Y}^\top \text{Softmax}\left(\beta \boldsymbol{Y} \boldsymbol{r}^{(t)}\right), \tag{5}$$

where $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)^\top$ denotes the memory bank, and $\boldsymbol{r}^{(t)}$ represents the input state pattern at time t. Further, we map each pattern to the associative space through projection matrices:

$$\boldsymbol{Z} = \boldsymbol{Y}\boldsymbol{W}_V \cdot \text{softmax}\left(\beta \boldsymbol{R}\boldsymbol{W}_Q \boldsymbol{W}_K^T \boldsymbol{Y}^T\right), \tag{6}$$

where $\boldsymbol{Z}$ denotes the result patterns, $\boldsymbol{R}$ represents the state patterns, and $\boldsymbol{Y}$ means the stored patterns. The projection matrices $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$, and $\boldsymbol{W}_V$ transform these patterns. The temperature β controls the sharpness of the softmax distribution. A high value of β indicates that each pattern is easier to retrieve individually. A low value of β indicates a higher likelihood of metastability, and the retrieved state will be a superposition of multiple stored patterns.

B.3 More Details about ATU

Update Process. The dynamic template update process is detailed in Algorithm 1. Specifically, div_scale is computed as the inverse similarity to quantify sample diversity. High-confidence templates reliably refresh the ST, whereas highly diverse samples are archived in the LT. Finally, we stochastically resample patterns and employ the Hopfield layer as a denoising reconstructor to synthesize the refined template $\mathbf{z}$.

For consistency with established paradigms in THOR (Sauer et al., 2019) and MambaEVT (Wang et al., 2025a), we adopt their settings for div_scale and ST/LT memory configurations. In addition, stochastic sampling is used as a test-time augmentation strategy to increase template diversity, reduce overfitting risk, and improve long-term robustness. Meanwhile, short-term template diversity is jointly regulated by update intervals and confidence scores. The update intervals control refresh frequency to capture varied target representations, whereas confidence scores filter out low-quality candidates and thus prevent memory pollution.

C Experiments

C.1 Implementation Details

In the training phase, we follow the settings of OSTrack (Ye et al., 2022) and use the AdamW optimizer. The learning rate is set to 4e-4, the weight decay is set to 1e-4, and the batch size is set to 12. We sample templates and search images in a 100-frame window, with 20,000 samples per epoch. In the testing phase, ST and LT are initialized with an initial static template at the beginning and dynamically updated with tracking results in subsequent phases. The default capacity for ST and LT is 5 and 10, respectively. All experiments are conducted on a computing server equipped with an NVIDIA GeForce RTX 4090 GPU. To ensure fair comparison, FPS is consistently evaluated on an NVIDIA GeForce RTX 2080 Ti GPU.

C.2 Ablation Study

• **Effectiveness of Proposed Components on FE108.** As shown in Table 8, we report the compo-

Table 8: component analysis on FE108 dataset and efficiency and parameter analysis.

#	ATU	AMAH	SR	PR	NPR	Params	FLOPs
1			60.4	91.6	65.0	92.52M	89.55G
2	✓		61.0	92.3	65.3	96.43M	92.48G
3		✓	62.0	93.2	67.6	97.25M	96.22G
4	✓	✓	**63.0**	**94.4**	**68.1**	101.16M	99.16G

nent analysis on FE108 with MAE pretrained weights. The results are consistent with the COESOT results in the main paper: both ATU and AMAH improve the performance, and their combination achieves the best results.

• **Efficiency and Parameter Analysis.** As shown in Table 8, our method remains efficient: the full model increases parameters by only 9.3% and FLOPs by 10.7% over the baseline, indicating that the performance gain is achieved with modest computational overhead.

• **Analysis on Associative Memory Augmented Hopfield.** As shown in Table 9, we analyze the detailed selection of Hopfield layers and preceding layers as stored patterns. The selection of stored patterns and preceding layer features in the Hopfield layer is crucial. We compare three schemes: *1).* $[3, 7, 11]$: $\mathcal{L} = \{3, 7, 11\}$, $\mathcal{P}_3 = \{0, 1, 2\}$, $\mathcal{P}_7 = \{4, 5, 6\}$, $\mathcal{P}_{11} = \{8, 9, 10\}$. *2).* $[5, 8, 11]^1$: $\mathcal{L} = \{5, 8, 11\}$, $\mathcal{P}_5 = \{1, 2, 3\}$, $\mathcal{P}_8 = \{4, 5, 6\}$, $\mathcal{P}_{11} = \{7, 8, 9\}$. *3).* $[5, 8, 11]^2$: $\mathcal{L} = \{5, 8, 11\}$, $\mathcal{P}_5 = \{1, 3\}$, $\mathcal{P}_8 = \{4, 6\}$, $\mathcal{P}_{11} = \{7, 9\}$. Finally, we choose $[5, 8, 11]^2$ as the setting of our Hopfield retrieval layer. The other two settings also improve the performance compared to the baseline.

Table 9: analysis of AMAH layers on COESOT.

Hopfield Layers	SR	PR	NPR
Baseline	63.2	77.1	76.1
$[3, 7, 11]$	63.4	77.5	76.6
$[5, 8, 11]^1$	63.6	77.9	76.7
$\mathbf{[5, 8, 11]^2}$	**63.9**	**78.4**	**77.1**