

Multimodal Neuromorphic Event-Frame Fusion in Domain-Generalized Vision Transformer for Dynamic Object Tracking

Taha Razzaq Asim Iqbal*
Tibbling Technologies
asim@tibbtech.com

Abstract

Object tracking is a fundamental task in computer vision with critical applications in autonomous driving, surveillance, and robotics. However, existing tracking solutions struggle in high-speed, real-time scenarios due to their reliance on conventional frame-based sensors optimized for low-frame-rate environments. Neuromorphic event-driven sensors offer a compute-efficient alternative by capturing continuous, asynchronous intensity changes, excelling in fast-motion detection. However, neuromorphic vision sensors have a lower spatial resolution, limiting their ability to capture fine textures crucial for object identification. Multimodal fusion techniques have been explored recently to leverage the complementary strengths of both frames+events modalities, incorporating optical flow estimation, motion compensation, and deformable convolutions. While these fusion models improve performance under rapid motion, they remain susceptible to domain shifts, leading to degradation when tested on out-of-distribution "unseen" target data. To address this challenge, we introduce an application of neuro-inspired, domain-generalized Winner-Take-All (WTA) mathematical layer that seamlessly integrates into the Vision Transformer (ViT) architecture. Our approach enhances domain invariance in object detection and tracking systems, particularly in environments with diverse lighting conditions and visual variations. We demonstrate that our technique significantly improves ViT performance for image classification, even when trained on a limited dataset. Additionally, we propose a multimodal AI framework that enables real-time object detection through the fusion of frame+event-based data.

1. Introduction

Object tracking is a fundamental challenge in computer vision (CV), with critical applications spanning autonomous driving, surveillance, and robotics [4, 5, 17], etc. Con-

temporary tracking solutions, however, face significant limitations when confronted with high-speed, real-time scenarios due to their optimization for low-frame-rate environments, a constraint imposed by conventional frame-based sensors [15]. This results in suboptimal tracking accuracy when dealing with rapid motion or dynamic scene change, posing a substantial challenge for real-world deployments. Neuromorphic event-based cameras, inspired by the biological architecture of the mammalian retina, offer a promising alternative to traditional frame-based sensors. These dynamic vision sensors (DVS) capture continuous, asynchronous intensity changes, excelling in the detection of fast-moving objects and efficient processing of high-speed events [11, 12]. This makes them particularly well-suited for real-time object tracking in dynamic environments. However, event-based cameras are not without limitations; their lower spatial resolution constrains their ability to capture fine textures crucial for precise object identification — a task at which frame-based sensors excel.

Recognizing the complementary strengths of event- and frame-based object tracking, significant research efforts have been directed towards developing efficient models and comprehensive datasets for high-speed motion and object tracking [3]. The optimal fusion of these disparate modalities necessitates precise spatio-temporal alignment across multiple frames. A commonly explored approach involves the estimation of optical flow fields between consecutive frames to account for motion [15]. Additional sophisticated techniques, including implicit motion compensation and deformable convolutions, have been investigated to enhance alignment accuracy [2, 3, 21]. These methods are typically integrated with fusion algorithms that employ either data-driven strategies or utilize multi-stream convolutional networks to amalgamate the frame and event data streams [10, 14]. While these fusion models generally demonstrate superior performance compared to traditional sensors in high-motion scenarios, their efficacy remains limited when confronted with complex,

*Corresponding author.

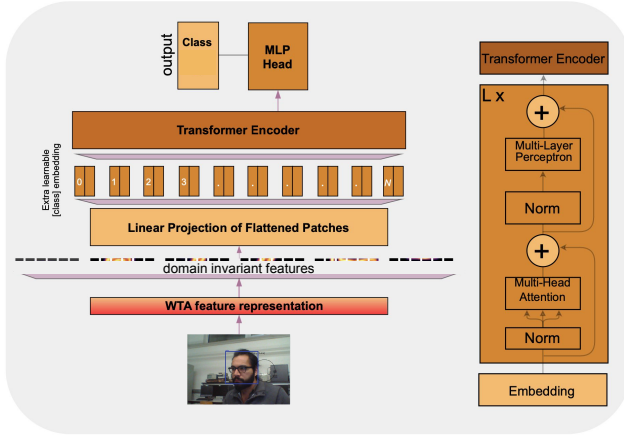


Figure 1. The image frame is passed through Vision Transformer architecture with our WTA layer to generate domain invariant features.

non-linear motion patterns [10]. Despite the significant advancements in multimodal data fusion for object tracking, a key challenge persists in ensuring domain invariance. Deep learning models often exhibit performance degradation when tested on out-of-distribution (OOD) data [7], compromising their reliability in diverse real-world conditions. In event-based fusion models, the inherent low spatial resolution of event data necessitates increased reliance on the frame-based component. This dependence can lead to reduced tracking accuracy when the frame-based model encounters unfamiliar or novel scenarios. To enhance robustness and extend the applicability of these systems, it is imperative to develop domain-invariant architectures capable of generalizing across varied data modalities. Such advancements would unlock the full potential of event- and frame-based fusion systems, significantly improving performance for critical real-world applications in autonomous driving, robotics, medicine, and beyond.

In this work, we test a neuro-inspired domain generalized Winner-Take-All (WTA) mathematical layer that seamlessly integrates into a Vision Transformer (ViT) architecture, as shown in Figure 1. This novel approach ([8]) directly addresses the critical challenge of domain invariance in object detection and tracking systems, particularly in environments with varied lighting conditions and visual characteristics.

2. Related Work

Transformer-based deep learning models have recently gained prominence in object detection and tracking tasks due to their ability to capture rich visual representations. By combining these models with event-based sensors, researchers have developed fusion algorithms that leverage

both the high temporal resolution of event-based sensors and the strong spatial representation of transformer models [11, 12, 20]. This fusion has significantly improved the detection and tracking of fast-moving objects, addressing the limitations of traditional frame-based models.

Early integration attempts focused on discretizing asynchronous event data for compatibility with conventional deep learning models. Specialized frame-and-event datasets like FE108 [12] and COESOT [16] were curated to facilitate benchmarking and experimentation. These datasets, recorded using DAVIS cameras [6], provide dense ground-truth annotations for both indoor and outdoor settings.

Despite promising results, early fusion efforts were constrained when event streams were sparse or incomplete. To address this, data-driven approaches employing multi-stream convolutional networks [2, 3, 12, 15] were introduced. These architectures can process both event and frame data more robustly, managing sparse event inputs while maintaining reliable object detection and tracking accuracy.

Recent advancements include RGB-event tracking [22, 23], which incorporates color information crucial for real-world applications like autonomous driving. A notable development is a transformer-based RGB-event tracking framework leveraging Vision Transformer (ViT) models [23]. This framework employs a mask modeling strategy for efficient cross-modal information utilization, achieving state-of-the-art performance on benchmark datasets.

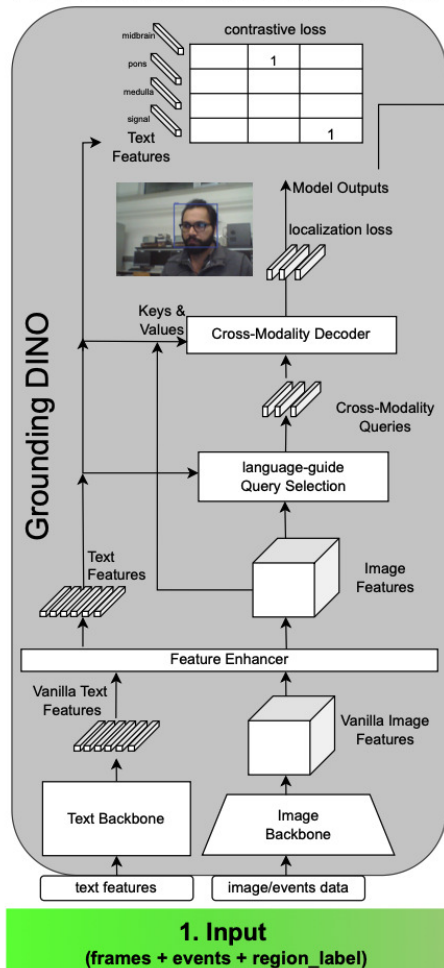
While these fusion algorithms [18–20, 22, 23] represent significant progress in real-time, high-motion object detection and tracking, several challenges remain. For instance, deep learning models often underperform when faced with covariate shifts between training and testing environments [7], such as differing lighting or weather conditions. This leads to deterioration in the model’s generalization capability and accuracy, as learned representations are biased toward specific training conditions. Consequently, enhancing domain invariance and adaptability across varying environments remains a critical area for further research and development.

3. Methodology

3.1. Domain adaptation with neuro-inspired frames/events fusion

Our approach takes a neuro-inspired, domain-generalized Winner-Take-All (WTA) mathematical layer that is designed for seamless integration into Vision Transformer

2. Feature classification



3. Feature segmentation

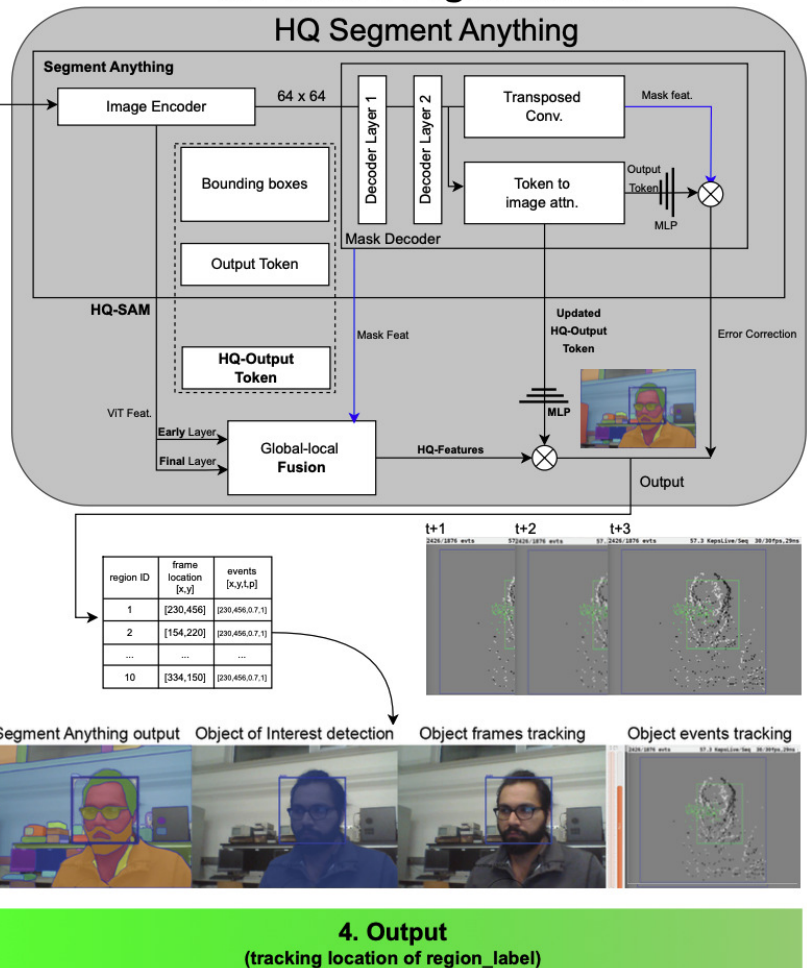


Figure 2. Multimodal AI framework: The input—comprising frames, events, or region labels—is first processed by a feature classifier that automatically identifies objects of interest using Grounding DINO, generating precise bounding boxes. Next, a pre-trained HQ-SAM model refines this localization through high-accuracy segmentation. Finally, joint fusion of localization of frames and events takes place for real-time object tracking.

(ViT) architectures, as shown in [Figure 1](#). This approach directly tackles the challenge of achieving domain invariance in object detection and tracking systems, particularly in environments with diverse lighting conditions and visual characteristics. The WTA layer representation, drawing inspiration from competitive mechanisms observed in biological neural networks, when added to the ViT architecture, significantly enhances the ViT’s ability to extract robust, domain-invariant features from both frames [8] and event data streams.

This layer mirrors neurobiological processes to enhance feature discrimination and reduce sensitivity to domain-specific noise, thereby improving the overall performance and generalization capabilities of the ViT

model ([8]). Complementing this is our proposed adaptive fusion mechanism, an advanced algorithm that dynamically weighs the contributions of frames and events data based on real-time environmental conditions. This domain adaptive approach ensures optimal performance across a diverse range of scenarios, from low-light conditions to rapidly changing environments.

Our innovation extends beyond individual components to encompass a comprehensive domain generalization framework. This framework governs the training and fine-tuning processes of the ViT architecture, maximizing its ability to generalize across domains and significantly reducing performance degradation on out-of-distribution data. This is particularly crucial for real-world applications

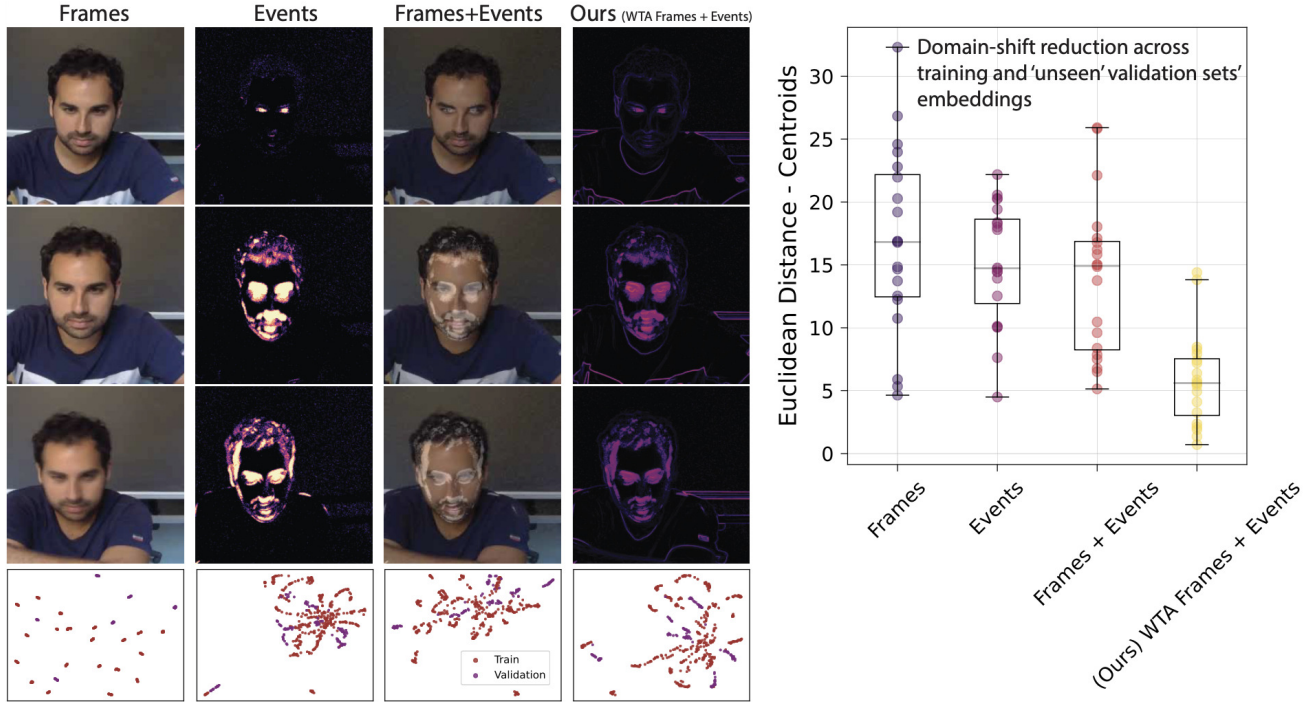


Figure 3. Comparison of representations from Frames, Events, Frames+Events and Our (WTA-based Frames + Events) representations (left). The bottom clusters show the training and ‘unseen’ validation sets’ UMAP embeddings with training data in brown and validation data in purple, here each data point in the cluster represents an image. The right box plot shows the euclidean distance between centroids of UMAP embeddings of each training (unique person) class with the validation (unique persons’) classes. Our neuro-inspired representation results in minimizing the overall domain shift between training and ‘unseen’ validation samples.

where deployment conditions may differ substantially from training data.

To mathematically implement the WTA technique, we process an input image (I) frame, where W , H , and C represent its width, height, and number of channels, respectively. To generate a WTA-driven domain invariant representation (I^G), we generate patches P of the image of size k .

$$I \in \mathbb{R}^{W \times H \times C} \quad (1)$$

$$P = \{p_{s_1}, p_{s_2}, \dots, p_{s_n}\} \quad (2)$$

$$\sigma_{p_s} = \left(\frac{1}{s^2} \sum_{i,j \in P,s} (k_{ij} - \mu_{p_s})^2 \right)^{1/2}, \quad (3)$$

where:

$$\mu_{p_s} = \frac{1}{s^2} \sum_{i,j \in P,s} k_{ij} \quad (4)$$

$$z = \max\{\sigma_{p_{s_1}}, \sigma_{p_{s_2}}, \dots, \sigma_{p_{s_n}}\} \quad (5)$$

$$I^G = \frac{\sigma_{p_s}}{z} \quad (6)$$

Each patch’s mean and standard deviation is computed and normalized by the max of standard deviation; which results in enhancing the features and suppressing the background noise. This mimics the WTA computation circuit motifs observed in the visual cortex where a variety of inhibitory neurons mediates the output of the excitatory neuron with varied inhibition mechanisms ([8])

3.2. Fusion of frames and events with Multimodal AI framework

In order to demonstrate the real-world capabilities of our proposed technique, using a frame-based webcam and a Dynamic Vision Sensor, we developed a frame/event-based object tracking framework that combines real-time motion capture with advanced AI models. For precise object detection, we incorporated state-of-the-art models such as Grounding DINO [13] for scene parsing and HQ-SAM [9] for semantic segmentation of each object-of-interest, as shown in Figure 2. This combination offers precise object detection and seamless real-time tracking, even under challenging conditions like rapid movement or poor light-

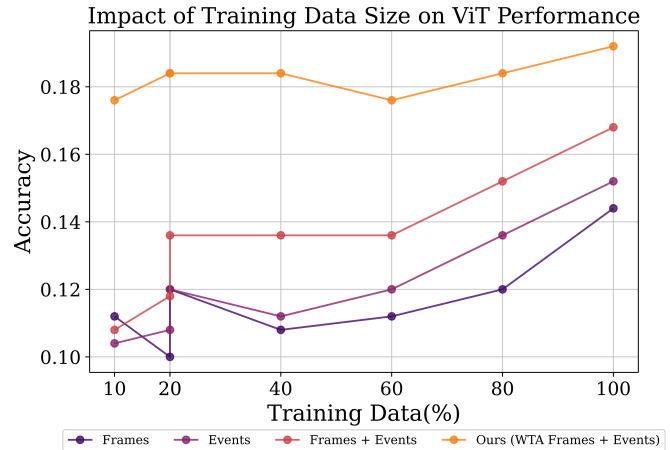
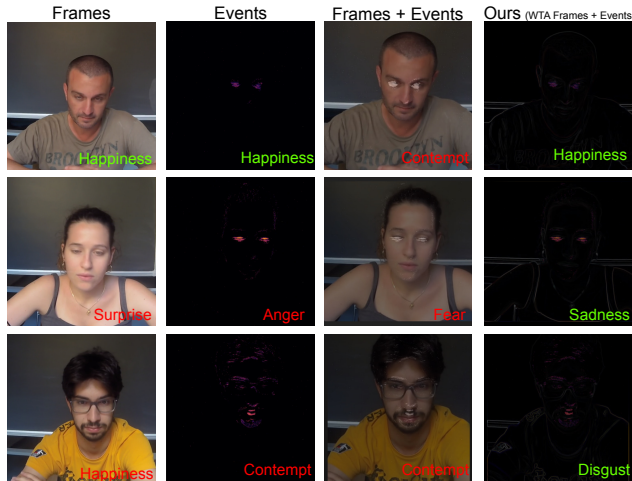


Figure 4. Qualitative and Quantitative Evaluation of ViT with WTA Integration. **Left:** Sample qualitative results for emotion classification using ViT across different representations—Frames, Events, Frames + Events, and our proposed WTA Frames + Events. The predicted emotion is displayed at the bottom right of each sample, with green indicating correct predictions and red indicating incorrect ones. **Right:** Quantitative performance comparison of ViT across the same representations, evaluated with progressively decreasing training data.

ing. The input frames are first passed through a feature classifier that labels the object(s)-of-interest automatically through Grounding DINO, resulting in the location of the object through a bounding box. The feature segmenter block captures the input from the feature classifier and segments the object(s)-of-interest with high accuracy through a pre-trained HQ-SAM model. The DVS events efficiently capture fine-grained motion information, enabling robust tracking in high-speed, dynamic environments. The DVS output is finally shared with the output of feature classifier and segmenter simultaneously.

4. Results & Discussion

4.1. Domain adaptation with neuro-inspired frames/events fusion on face datasets

We quantified the performance of the Winner-Take-All (WTA) technique by maximizing the domain shift robustness or minimizing the domain shift present between training (seen) and validation (unseen) data [1] for AI models, as shown in Figure 3. This open-source dataset, NEFER, contains multimodal frames/events data of 26 subjects across varying facial expressions (with diverse emotions), making it ideal data to test the performance of our domain adaptive technique.

To test the domain adaptive capability, after passing the frames/events through the WTA mathematical image processing layer, we generated the UMAP embeddings of the entire training (20 subjects x 21 expression x 62 frames/events) with the entire validation set (6 subjects x 21 expressions x 62 frames/events). We also computed the

same for only frames, events and frames+events (without the WTA technique). The bottom left panel in Figure 3 demonstrates our results and we further quantified the Euclidean distance between the centroids of training and validation clusters for each subject across the entire training and unseen validation sets. Our technique shows minimum domain shift (Figure 3; right plot) on the entire dataset as compared to the baseline representations (frames, events, frames/events without our WTA technique).

To further evaluate the effectiveness of our deployed WTA layer in object detection and emotion classification, we fine-tuned ViT on a subset of the training data and assessed its performance on a consistent test set across four representations: frames only, event only, frames+event, and our proposed WTA frames+events representation. We progressively reduced the training dataset size and observed that the WTA representation of frames+events achieved the highest overall accuracy, with only a slight decline even as the training samples were significantly reduced, as shown in Figure 4 (right plot). In contrast, the other representations exhibited lower accuracy and experienced a substantial drop in performance as the training data decreased. Additionally, Figure 4 presents the quantitative results of the model’s performance across these four representations, for 3 different emotions class labels.

4.2. Fusion of frames and events with multimodal AI framework on real-world tracking

To evaluate our frame/event-based object tracking framework, we assess its performance on a face-tracking task under multiple motion conditions (Figure 5), demonstrat-

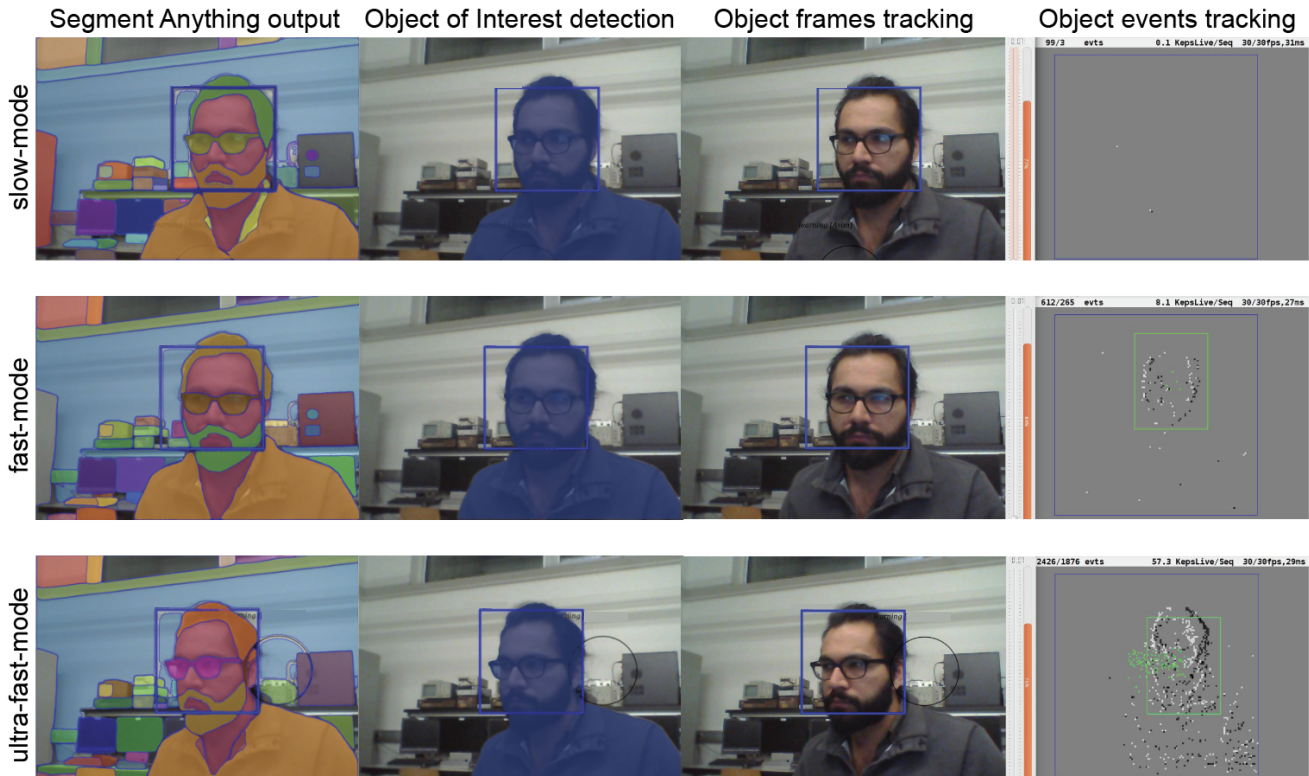


Figure 5. Joint segmentation and tracking of object of interest through our multimodal AI framework with (slow, fast and ultra-fast) multiple modes to track object (with face shortlisted as example here) by a high-performing Segment Anything model. Frames and events are processed through Grounding DINO for region localization, followed by segmentation using HQ-SAM. The top row illustrates the slow mode, where no events are detected, and object detection relies solely on frames. The middle row represents the fast mode, where despite sparse events being generated, consistent tracking is ensured. The bottom row depicts the ultra-fast mode, where both events and frames contribute to object detection for enhanced performance.

ing the integration of HQ-SAM, Grounding DINO, and DVS-based motion event tracking across slow, fast, and ultra-fast settings. Using a frame-based webcam and a DVS sensor, we perform real-time motion detection, where events are processed in real-time alongside every 100th frame captured by the webcam. The frames and event data are fed into Grounding DINO to localize the region of interest, which is subsequently segmented using HQ-SAM. Real-time object tracking is achieved through the joint fusion of frame-based and event-based localization outputs.

In the slow mode (top row; [Figure 5](#)), no events are detected, but our model successfully identifies the face using frames alone. During fast mode (middle row; [Figure 5](#)), sparse events are generated, yet our framework maintains continuous tracking and detection. In ultra-fast mode (bottom row; [Figure 5](#)), the event camera captures the face’s trajectory with high temporal precision, while the frame-based module performs precise object detection. The fusion algorithm shares real-time bounding box localization

$[x, y]$ of the object of interest and corresponding dynamic events $[x, y, t, p]$. In our experiments, the frame-rate is sampled to match the sampling rate of the events from the dynamic vision sensor. This event-frame fusion approach ensures consistent performance across various motion scenarios, showcasing our framework’s versatility and effectiveness in dynamic environments.

5. Conclusion

In this study, we introduce a neuro-inspired, domain-generalized Winner-Take-All (WTA) mathematical layer that seamlessly integrates into the Vision Transformer (ViT) architecture. To assess its effectiveness, we evaluate the model on an object and emotion detection task, demonstrating that our proposed WTA representation of frames + events consistently outperforms conventional approaches. Our findings highlight the ability of the WTA layer to enhance model robustness and domain invariance, particularly in challenging environments with varying lighting conditions and visual characteristics. Furthermore, our results indicate that even with a significantly reduced train-

ing dataset, the WTA-integrated ViT maintains high performance, underscoring its potential for real-world applications in resource-constrained settings. This work paves the way for further advancements in neuromorphic-inspired AI architectures, enabling more reliable and efficient multimodal object tracking and recognition systems.

References

- [1] Lorenzo Berlincioni, Luca Cultrera, Chiara Albisani, Lisa Cresti, Andrea Leonardo, Sara Picchioni, Federico Becattini, and Alberto Del Bimbo. Neuromorphic event-based facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4109–4119, 2023. 5
- [2] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 1, 2
- [3] Jianing Deng, Li Wang, Shiliang Pu, and Cheng Zhuo. Spatio-temporal deformable convolution for compressed video quality enhancement. In *Proceedings of the AAAI conference on artificial intelligence*, pages 10696–10703, 2020. 1, 2
- [4] Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens, Lucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. Tap-vid: A benchmark for tracking any point in a video. *Advances in Neural Information Processing Systems*, 35:13610–13626, 2022. 1
- [5] Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. Tapir: Tracking any point with per-frame initialization and temporal refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10061–10072, 2023. 1
- [6] 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. 2
- [7] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016. 2
- [8] Asim Iqbal, Hassan Mahmood, Greg J Stuart, Gord Fishell, and Suraj Honnuraiah. Biologically realistic computational primitives of neocortex implemented on neuromorphic hardware improve vision transformer performance. *bioRxiv*, pages 2024–10, 2024. 2, 3, 4
- [9] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36: 29914–29934, 2023. 4
- [10] Min Seok Lee, Ye Jun Kim, Jae Hyung Jung, and Chan Gook Park. Fusion of events and frames using 8-dof warping model for robust feature tracking. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 834–840. IEEE, 2023. 1, 2
- [11] Martin Litzemberger, Christoph Posch, Daniel Bauer, Ahmed Nabil Belbachir, Peter Schon, Bernhard Kohn, and Heinrich Garn. Embedded vision system for real-time object tracking using an asynchronous transient vision sensor. In *2006 IEEE 12th Digital Signal Processing Workshop & 4th IEEE Signal Processing Education Workshop*, pages 173–178. IEEE, 2006. 1, 2
- [12] Qianhui Liu, Dong Xing, Huajin Tang, De Ma, and Gang Pan. Event-based action recognition using motion information and spiking neural networks. In *IJCAI*, pages 1743–1749, 2021. 1, 2
- [13] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 4
- [14] Genady Paikin, Yotam Ater, Roy Shaul, and Evgeny Soloveichik. Efi-net: Video frame interpolation from fusion of events and frames. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1291–1301, 2021. 1
- [15] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6626–6634, 2018. 1, 2
- [16] 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. 2
- [17] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking everything everywhere all at once. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19795–19806, 2023. 1
- [18] 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*, pages 19248–19257, 2024. 2
- [19] 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*, pages 13043–13052, 2021.
- [20] 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*, pages 8801–8810, 2022. 2
- [21] 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, pages 9781–9790, 2023. [1](#)

- [22] Zhuyun Zhou, Zongwei Wu, Rémi Boutteau, Fan Yang, Cédric Démonceaux, and Dominique Ginjac. Rgb-event fusion for moving object detection in autonomous driving. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7808–7815. IEEE, 2023. [2](#)
- [23] 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*, pages 22045–22055, 2023. [2](#)