

# SIS-Challenge: Event-based Spatio-temporal Instance Segmentation Challenge at the CVPR 2025 Event-based Vision Workshop

Friedhelm Hamann<sup>1</sup>, Emil Mededovic<sup>2</sup>, Fabian Gülhan<sup>2</sup>, Yuli Wu<sup>2</sup>,  
Johannes Stegmaier<sup>2</sup>, Jing He<sup>3</sup>, Yiqing Wang<sup>3</sup>, Kexin Zhang<sup>3</sup>,  
Lingling Li<sup>3</sup>, Licheng Jiao<sup>3</sup>, Mengru Ma<sup>3</sup>, Hongxiang Huang<sup>4</sup>,  
Yuhao Yan<sup>5</sup>, Hongwei Ren<sup>4</sup>, Xiaopeng Lin<sup>4</sup>, Yulong Huang<sup>4</sup>,  
Bojun Cheng<sup>4</sup>, Se Hyun Lee<sup>6</sup>, Gyu Sung Ham<sup>6</sup>, Kanghan Oh<sup>6</sup>,  
Gi Hyun Lim<sup>6</sup>, Boxuan Yang<sup>7</sup>, Bowen Du<sup>7</sup>, and Guillermo Gallego<sup>1</sup>

<sup>1</sup> TU Berlin, SCIOI, ECDF, <sup>2</sup> RWTH Aachen, <sup>3</sup> Xidian University,

<sup>4</sup> Hong Kong University of Science and Technology,

<sup>5</sup> Sun Yat-sen University, <sup>6</sup> Wonkwang University, <sup>7</sup> Tongji University.

## Abstract

We present an overview of the Spatio-temporal Instance Segmentation (SIS) challenge held in conjunction with the CVPR 2025 Event-based Vision Workshop. The task is to predict accurate pixel-level segmentation masks of defined object classes from spatio-temporally aligned event camera and grayscale camera data. We provide an overview of the task, dataset, challenge details and results. Furthermore, we describe the methods used by the top-5 ranking teams in the challenge. More resources and code of the participants' methods are available here: [https://github.com/tub-rip/MouseSIS/blob/main/docs/challenge\\_results.md](https://github.com/tub-rip/MouseSIS/blob/main/docs/challenge_results.md)

## 1. Introduction

With the rapid evolution of computer vision applications in robotics, autonomous systems, and biological research, the ability to accurately segment and track multiple objects over time has become increasingly important. Traditional frame-based cameras, while widely adopted, face fundamental limitations when dealing with challenging visual conditions such as high-speed motion, varying illumination, and low-light environments. These limitations are particularly pronounced in applications requiring real-time performance and high temporal precision, such as tracking tasks applicable to many problems, for example, behavioral analysis in neuroscience research and wildlife monitoring.

Event cameras, also known as Dynamic Vision Sensors (DVS) [5, 13], offer a compelling alternative to conventional frame-based sensors. Unlike traditional cameras that capture full images at a fixed rate, event cameras asyn-

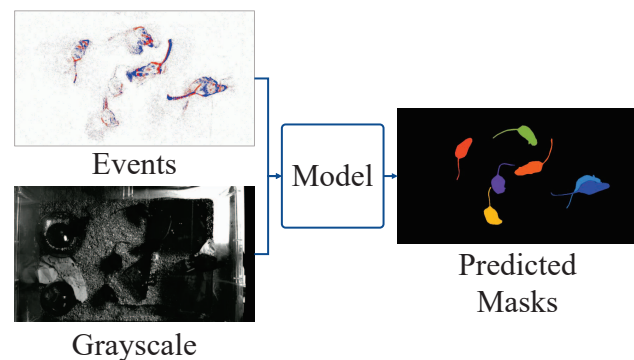


Figure 1. The Spatio-Temporal Instance Segmentation Challenge. Participants build models that take pixel-level aligned grayscale frames and events as input and predict accurate pixel-level segmentation masks and identifiers of all objects of class “mouse”.

chronously detect pixel-level brightness changes, producing a sparse stream of events only when and where changes occur in the scene. This unique sensing paradigm provides several advantages: microsecond temporal resolution, high dynamic range (>120 dB), low power consumption, and minimal motion blur [6]. These characteristics make event cameras particularly well-suited for tracking fast-moving objects under challenging lighting conditions, where frame-based approaches often fail.

Despite these advantages, the adoption of event-based vision for complex tracking tasks has been limited by the lack of annotated datasets that support fine-grained, multi-object tracking at the pixel level. Although significant progress has been made in frame-based video instance segmentation [21, 23], the event-based vision community has

focused primarily on simpler tasks, such as single-object bounding-box tracking. This gap in available resources has hindered the development of sophisticated event-based tracking algorithms that could fully leverage the unique properties of event cameras.

To address this critical need, we present the Spatio-temporal Instance Segmentation (SIS) Challenge, organized as part of the 2025 CVPR Event-based Vision Workshop<sup>1</sup>. This challenge is based on the MouseSIS dataset [8], which provides aligned event and frame data with pixel-accurate instance segmentation masks for multiple freely moving mice. The dataset includes 33 video sequences with an average duration of approximately 20 seconds, recorded using a beamsplitter system that ensures pixel-level alignment between frames and events. The sequences contain challenging scenarios, including uneven illumination, occlusions, and complex interactions between multiple targets.

The 2025 SIS Challenge explores algorithmic potentials for multi-object mask-level tracking using event data. Unlike traditional video instance segmentation tasks that operate on images, the SIS Challenge addresses the unique opportunities presented by the quasi-continuous nature of event streams. Participants were tasked with developing methods that could accurately segment and track multiple mouse instances throughout entire sequences while maintaining consistent temporal identities (Fig. 1). The Challenge ran from February to May 2025, attracting 63 participants with 14 teams submitting results to the leader board.

This paper summarizes the approaches and findings from the top-performing teams in the Challenge. The results demonstrate that event-based approaches can achieve competitive performance in complex multi-object tracking scenarios, with the winning method achieving a Higher Order Tracking Accuracy (HOTA) score of 0.62. Hence, the SIS Challenge and this accompanying summary contribute to the advancement of the field of event-based computer vision, showcasing the potential of event cameras for complex scene understanding tasks. By providing a benchmark for event-based spatio-temporal instance segmentation, we aim to inspire future research in further developing robust tracking algorithms that can operate effectively under challenging visual conditions where traditional cameras struggle.

## 2. Spatio-Temporal Instance Segmentation Challenge

### 2.1. Introduction of the SIS Dataset

The SIS Challenge is based on the MouseSIS dataset [8], a benchmark for multi-object tracking and segmentation using synchronized event and frame data. The dataset captures freely moving mice in laboratory settings using a

specialized hardware setup that ensures pixel-level alignment between neuromorphic event cameras and conventional grayscale cameras. The MouseSIS dataset comprises 33 sequences, each approximately 20 s in duration: around 600 frames at 30 Hz and aligned event data. The sequences feature varying numbers of mice (1–6 subjects) engaged in natural behaviors under different lighting conditions, including challenging scenarios with occlusions, rapid movements, and uneven illumination. The dataset follows a YouTubeVIS-style annotation format, providing instance-level segmentation masks with consistent identifiers (IDs) throughout each sequence. Data is organized into predefined train, validation, and test splits, with sequences distributed to ensure balanced difficulty across splits.

### 2.2. Task Description

The Challenge requires participants to develop algorithms for spatio-temporal instance segmentation of mice from synchronized event and frame data (Fig. 1). Specifically:

1. **Input:** Participants receive pixel-aligned event streams and grayscale frames for each test sequence. Event data is provided as raw events  $(x, y, t, p)$  where  $(x, y)$  are pixel coordinates,  $t$  is the timestamp, and  $p$  is polarity.
2. **Output:** Methods must produce temporally accurate and consistent pixel-level instance segmentation masks and object IDs for all mice in each sequence.
3. **Evaluation:** Following the MouseSIS evaluation protocol, methods are assessed using multiple metrics, including HOTA [14], Multiple Object Tracking Accuracy (MOTA) [1], and IDF1 [18] scores, which jointly evaluate segmentation quality and temporal consistency.

For the Challenge, participants process six test sequences: 10, 16, 22, 26, 28 & 32. Sequences 1 and 7 from the original test set [8] were excluded to maintain evaluation integrity. Submissions consist of JavaScript Object Notation (JSON) files containing predicted segmentation masks in Run-Length Encoding (RLE) format with associated instance IDs and confidence scores.

### 2.3. Data Loading and Training Pipeline

To facilitate participation and ensure reproducibility, the Challenge provides a comprehensive codebase with standardized data loading and training pipelines:

**Data Access.** Participants can download the MouseSIS dataset from the provided Google Drive repository, organized in HDF5 format with separate files for each sequence. Each HDF5 file contains synchronized frames and events with precise temporal alignment information.

**Preprocessing Pipeline.** The codebase includes utilities:

1. Loading and synchronizing event and frame data from HDF5 files.
2. Converting raw events to various representations (e.g., event frames, voxel grids).

<sup>1</sup><https://tub-rip.github.io/eventvision2025/>

3. Handling the YouTubeVIS-style annotations with proper sequence-instance mapping.

**Baseline Implementation.** A complete baseline method, *ModelMixSort* [8], that combines YOLOv8 object detection and Segment Anything Model (SAM)-based segmentation with XMem-based tracking, is provided. It demonstrates:

1. Multi-modal fusion of events and frames.
2. Integration with popular deep learning frameworks (e.g., PyTorch).
3. Standard training procedures with configurable hyperparameters.
4. Inference scripts for generating Challenge-compliant JSON outputs.

The pipeline supports flexible experimentation while maintaining standardized evaluation procedures, enabling a fair comparison of different approaches.

### 3. Challenge Results

This section summarizes the results of the top-5 teams in the Challenge ranking, showing an increase of up to 42% compared to the baseline method *ModelMixSort* [8]. Most teams follow a similar tracking-by-detection approach as *ModelMixSort*, improving this modular method by integrating and fine-tuning the latest detection and segmentation methods in the literature. Technical details of all methods can be found in Sec. 5. In summary, the centralized evaluation and modular baseline method provide easy access, also for non-event-vision practitioners, to the topic of event-based tracking. This low entry barrier allows participants to integrate the latest advances in foundation models and explore the advantages of event-based cameras.

Team Name	HOTA $\uparrow$	MOTA $\uparrow$	IDF1 $\uparrow$
1. emilmed	<b>0.62</b>	<b>0.72</b>	<b>0.83</b>
2. enidx	0.57	0.69	0.75
3. mysterypeople	0.54	0.59	0.67
4. shlee	0.54	0.61	0.68
5. vivien	0.54	0.54	0.67
ModelMixSort (baseline) [8]	0.43	0.45	0.50

Table 1. Top-5 results of the CVPR 2025 Spatio-temporal Instance Segmentation (SIS) Challenge at the Event-based Vision Workshop. The HOTA score determines the overall ranking. Bold values indicate the best results per metric.

### 4. Conclusion

We presented the results of the SIS Challenge held in conjunction with the CVPR’25 Event-based Vision Workshop. Progress in event-based vision in general, and more specifically in event-based tracking, is lagging behind conventional vision in terms of easily accessible evaluation platforms. The MouseSIS dataset and the SIS Challenge provide steps towards closing such a gap. This report provides

an overview of the MouseSIS dataset, the challenge, and technical details of the top-5 methods. The solutions of the participants show creative integration of existing frame-based methods and optimizations, which improve accuracy by  $\approx 42\%$  compared to the baseline method. We believe that this benchmark, which includes an accessible baseline method and centralized evaluation, significantly lowers the entry barrier to event-based tracking and helps foster future developments in this topic.

## 5. Challenge Teams and Methods

### 5.1. Team 1: emilmed

#### 5.1.1. Description

The tracking pipeline of the winning team (Fig. 2) draws inspiration from [19], which emphasizes that carefully selected design choices within a classical tracking-by-detection paradigm can achieve competitive performance, and also highlights the critical role of domain adaptation.

Pretrained Convolutional Neural Networks (CNNs) [15], like those in YOLOv8, rely on filters that assume a certain dynamic range in input intensities. Low-contrast inputs violate this assumption, resulting in weak activations and poorly separated feature maps. Histogram equalization improves contrast and reduces the distributional mismatch with COCO-pretrained detectors [11], enhancing the responsiveness of early convolutional filters. This constitutes a lightweight form of domain adaptation that aligns low-level input statistics. They retrain the detection model on these equalized images and finetune SAM-Large [12] using Low-Rank Adaptation (LoRA) [10], enabling efficient and scalable domain-specific adaptation.

For object association, the team adopts a more classical strategy based on bounding boxes. This approach offers faster and more stable motion prediction compared to mask-level tracking, which can be noisy and computationally demanding in complex scenes. In the following, the key components of the tracking pipeline are elaborated.

**Object Detectors.** For frame-based input, the team applies histogram equalization during training to mitigate lighting inconsistencies and retrains the YOLOv8 [11] detector model using the equalized frames. For event-based input, they use the provided baseline detection model applied to E2VID-reconstructed frames [17].

**Segmentation.** The team finetunes SAM [12] by varying only the decoder, leaving the image encoder untouched (i.e., frozen). To enable efficient adaptation, they inject LoRA [10] into the encoder’s attention projections, modifying weight matrices as follows:

$$W \leftarrow W + AB, \quad A \in \mathbb{R}^{d \times r}, \quad B \in \mathbb{R}^{r \times d}, \quad r \ll d,$$

where  $A, B$  are trainable and initialized using the method proposed by He et al. [9]. This reduces the trainable parameters while enabling effective task-specific adaptation.

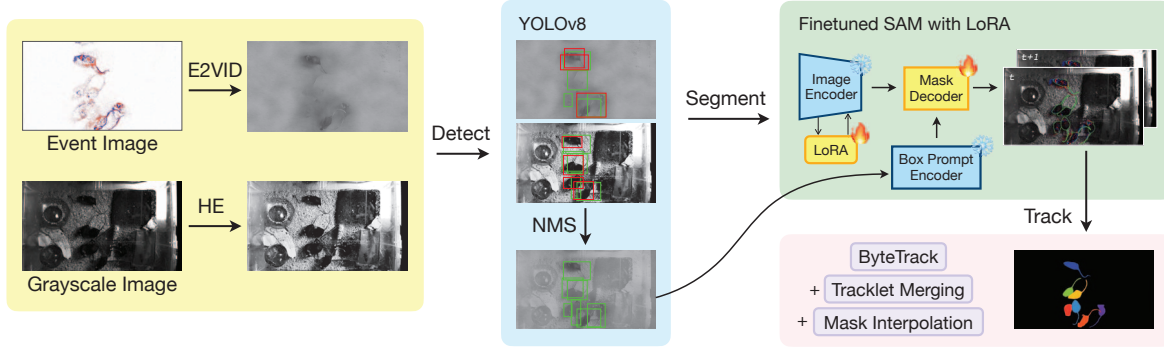


Figure 2. *Team 1*. Overview of the complete tracking pipeline. First, events are reconstructed using E2VID [17], while grayscale images are histogram-equalized (HE). These inputs are then passed to individually trained YOLOv8 detectors [11]. To eliminate duplicate detections, non-maximum suppression (NMS) is applied based on Intersection over Union (IoU) and detection confidence scores. The resulting bounding boxes are used to prompt a finetuned SAM model [12]. Depending on whether the detection originated from the reconstructed or histogram-equalized image, the corresponding input is fed into the SAM encoder for mask prediction. Tracking is performed using the ByteTrack paradigm [24], relying solely on bounding boxes. Finally, the tracking output is refined through a greedy merging of fragmented tracks, and missing masks are interpolated to produce the final results.

**Post-Processing.** The team refines the predicted segmentation masks using morphological operations inspired by the SAM2 pipeline [16]. Specifically, they apply dilation with a  $3 \times 3$  rectangular kernel for 4 iterations, followed by erosion for 3 iterations with the same structuring element. This operation smooths mask boundaries and fills small internal holes, resulting in more coherent and visually consistent segmentations across frames.

**Tracker.** The method employs ByteTrack [25] for multi-object tracking and performs data association using a cost matrix that combines Intersection-over-Union (IoU) similarity with detection confidence scores. Specifically, it weighs IoU similarities by detection confidence before converting them into a fused cost matrix. Appearance-based features were omitted due to the high visual similarity between rodents, which renders embedding-based metrics unreliable.

**Track Merging.** The greedy tracklet merging algorithm used in the pipeline of Fig. 2 is based on temporal proximity and spatial overlap (see Algorithm 1). The procedure considers pairs with small frame gaps and sufficient IoU between their bounding boxes. Pairs are greedily merged according to descending IoU.

**Track Interpolation.** To handle missing detections within a track, the method performs linear interpolation of object centroids at time point  $t$  and spatially shifts the segmentation masks. Let  $M_1 : \mathbb{Z}^2 \rightarrow \{0, 1\}$  be a binary mask defined on a 2D pixel grid. If there is a gap in the object’s tracked trajectory, it is spatially interpolated by translating its centroid. The team calculates the centroids of the detections at the gap borders using image moments [7]:

$$(c_x^{(\tau)}, c_y^{(\tau)}) = (M_{10}^{(\tau)} / M_{00}^{(\tau)}, M_{01}^{(\tau)} / M_{00}^{(\tau)}),$$

for  $\tau = \{t_1, t_2\}$ . The interpolated centroid is:

$$\begin{aligned} c_x^{\text{interp}} &= (1 - \alpha)c_x^{(t_1)} + \alpha c_x^{(t_2)} \\ c_y^{\text{interp}} &= (1 - \alpha)c_y^{(t_1)} + \alpha c_y^{(t_2)} \end{aligned} \quad \text{with} \quad \alpha = \frac{t - t_1}{t_2 - t_1}.$$

Then the required translation is computed:

$$\Delta x = c_x^{\text{interp}} - c_x^{(t_1)}, \quad \Delta y = c_y^{\text{interp}} - c_y^{(t_1)}.$$

The method constructs a 2D affine transformation matrix to perform this translation:

$$T = \begin{bmatrix} 1 & 0 & \Delta x \\ 0 & 1 & \Delta y \\ 0 & 0 & 1 \end{bmatrix}$$

This matrix is applied to  $M_{t_1}$  via image warping, yielding the interpolated mask  $M_t$  at time  $t$ , aligned with the linearly estimated object trajectory between frames  $t_1$  and  $t_2$ :

$$M_t(x, y) = M_1(T^{-1}[x, y, 1]^\top)$$

This procedure is repeated for each gap frame to reconstruct intermediate segmentation masks.

### 5.1.2. Implementation Details

**Object Detector.** The team retrained the YOLOv8m object detector [11] on the provided MouseSIS dataset [8]. The model was initialized with pretrained weights and trained for 100 epochs with a batch size of 16. Input images were resized to 640 pixels on the longest side while preserving aspect ratio. Optimization was performed using an SGD optimizer (with momentum 0.937 and weight decay 0.0005), with an initial learning rate of 0.01 and a 3-epoch warm-up phase, as per the default setting. To improve generalization,

---

**Algorithm 1** (*Team 1*). Greedy Tracklet Merging Based on Temporal and Spatial Continuity

---

**Require:** Tracklets  $\mathcal{T} = \{T_i\}$  with  $(s_i, e_i, B_i^{\text{first}}, B_i^{\text{last}})$

**Output:** Merged tracks

```
1: Initialize empty list of merge candidates  $\mathcal{C}$ 
2: for all pairs  $(T_i, T_j)$  with  $i \neq j$  do
3:   Compute frame gap  $\Delta_{ij} = s_j - e_i$ 
4:   if  $\Delta_{\min} < \Delta_{ij} \leq \Delta_{\max}$  then
5:     Compute  $\text{IoU}_{ij} = \text{IoU}(B_i^{\text{last}}, B_j^{\text{first}})$ 
6:     if  $\text{IoU}_{ij} \geq \theta$  then
7:       Add  $(i, j)$  to candidate list  $\mathcal{C}$ 
8:     end if
9:   end if
10: end for
11: Sort  $\mathcal{C}$  by descending IoU
12: for each pair  $(i, j) \in \mathcal{C}$  do
13:   if  $T_i$  and  $T_j$  are not already merged then
14:     Merge  $T_i$  and  $T_j$  into a single track
15:   end if
16: end for
```

---

they applied several augmentations during training, including Mosaic augmentation, horizontal flipping with a probability of 0.5, random translations ( $\pm 10\%$ ), scaling ( $\pm 50\%$ ) and random erasing with a probability of 0.4.

**Segmentation.** The team finetuned SAM [12] using the MouseSIS dataset [8]. The model was initialized with the official sam\_vit\_l\_0b3195 checkpoint. LoRA modules with rank  $r = 16$  were added to the query-key-value projections of selected self-attention layers in the image encoder. During training, only the LoRA modules and the mask decoder were updated, while the rest of the model remained frozen. Each image was resized to 1024 px on the longest side, normalized to  $[0, 1]$ , and padded to  $1024 \times 1024$  px. Training was carried out using the Adam optimizer with a learning rate of  $10^{-5}$  for up to 500 epochs, with gradient accumulation every 8 steps to simulate a larger batch size. The loss combined binary cross-entropy and Dice loss (weighted 0.005). The team manually applied early stopping based on validation loss.

**Tracking.** During tracking, the team performed non-maximum suppression between bounding boxes predicted from the E2VID reconstructions and frames. The finetuned SAM model was then applied to the E2VID input, as it better matches the training domain compared to the off-the-shelf model. They adjusted the tracker’s association parameters to suit the task, as follows: `track_high_thresh` = 0.6, `track_low_thresh` = 0.1, `match_thresh` = 0.8, `new_track_thresh` = 0.7, `track_buffer` = 60. They set the temporal association bounds to  $\Delta_{\min} = -15$  and  $\Delta_{\max} = 15$ , and used an IoU threshold  $\theta = 0.1$  for linking tracklets.

All training and inference runs have been performed on

a single NVIDIA RTX 3090 graphics card.

### 5.1.3. Results

The proposed pipeline achieved first place on the MouseSIS challenge. The results are reported in Tab. 1. The team outperformed other proposed solutions in all tracking metrics. However, the computation time is relatively high (2 seconds per sample, excluding E2VID runtime, as images are reconstructed in advance) and could benefit from model distillation to improve efficiency.

## 5.2. Team 2: enidx

### 5.2.1. Description

The team proposes a series of enhancements based on ModelMixSort, a tracking-by-detection approach in [2], to improve tracking and segmentation performance. Specifically, they replace YOLOv8 with the more powerful YOLOv12 for object detection, and substitute SAM2 with an upgraded version of SAM to enhance feature extraction capabilities. For preprocessing, they apply contrast enhancement to grayscale frames, and during inference, they employ test-time augmentations such as image rotation and flipping to improve model robustness and generalization. As a result, their method achieves great performance on the test set, ranking second overall (Tab. 1): HOTA=0.569, MOTA=0.688 and IDF1=0.748.

### 5.2.2. Implementation Details

**Input Process.** The MouseSIS dataset captures the activities of multiple mice in complex environments using synchronized event and frame cameras [8]. It provides high-quality instance segmentation masks, bounding boxes, and identity annotations for each mouse across video frames. The dataset includes numerous challenging scenarios involving frequent occlusions and interactions, making it well-suited for evaluating robust multi-object tracking and segmentation methods.

To improve the performance of downstream object detection and instance segmentation models, the team applies contrast enhancement as a standardized preprocessing step for all input grayscale frames. This design addresses the issue of weak feature visibility and blurred object boundaries, which are common in low-light or low-contrast scenes, and often degrade the discriminative capability of deep-learning models. Specifically, they adopt the Contrast Limited Adaptive Histogram Equalization (CLAHE) method with parameters `clipLimit` 2.0 and `tileGridSize` (8,8) to enhance the local contrast of each frame. By redistributing the grayscale values within localized regions, CLAHE expands the dynamic range and improves the visibility of fine details and textures that are otherwise hard to detect.

Contrast enhancement is applied consistently across the training, validation, and testing sets to ensure feature distribution alignment throughout the entire pipeline. This con-

sistency helps the model learn stable representations and mitigates performance degradation due to distribution shift.

**Boxes Detection.** To achieve accurate detection of mice, the team adopts YOLOv12 [20], a recent advancement over YOLOv8 with improved detection performance and feature extraction capability. Specifically, they use the large-scale YOLOv12-X variant to take advantage of its enhanced representational power. Given the multimodal nature of the data, they train two separate YOLOv12-X detectors: one for the event data and another for the grayscale frames. For the event modality, they first reconstruct frames from raw event streams using E2VID [17].

The dataset was partitioned by sequence, with frames 400 to 757 from sequence 33 designated as the validation set, while all remaining frames from the validation dataset were used for training. During training, all input images were uniformly resized to 640×640 pixels. The model was initialized with YOLOv12-X weights pretrained on COCO to accelerate convergence and enhance generalization performance. A batch size of 8 was employed during training, and each detector was trained for 80 epochs on its corresponding dataset. All other training configurations, including the choice of optimizer, learning rate scheduling policy, strictly followed the default settings provided by the official YOLO implementation to ensure consistency and reproducibility.

**Segmentation and Tracking.** After obtaining high-quality detection boxes from the YOLOv12-X, the team further employed SAM2 [16] for fine-grained instance segmentation. Specifically, they selected the more advanced SAM2.1\_hiera\_large, which demonstrates significantly improved segmentation performance compared to the original SAM [12]. The detection boxes produced by YOLOv12-X were used as prompts to guide SAM2 in generating corresponding segmentation masks. For the video object segmentation model, they used XMem [4], set to follow [8].

In order to make SAM2 better adapt to the specific appearance and pose variations of the mice in the MouseSIS dataset, the team fine-tuned the model according to the officially published pre-training weights. The fine-tuned dataset consists of MouseSIS training and validation datasets of grayscale frames.

The fine-tuning was performed using four NVIDIA GeForce RTX 4090 GPUs, with the batch size set to 1, the maximum number of objects processed per image capped at 6, and the resolution of all input images set to 1280 px, taking into account the GPU memory limitations and the model’s requirement for multi-object segmentation. Base learning rate was set to  $5 \cdot 10^{-6}$ , and the model was trained for a total of 40 epochs, with the rest of the training parameters following the default settings recommended in the SAM2 open-source implementation.

**Test-Time Augmentation.** In order to improve the stability and accuracy of the detection frame and thus enhance the segmentation of the SAM2, the team designed and implemented a test-time enhancement strategy. The strategy works by applying geometric transformations to the input images, such as horizontal flip, scaling, and small-angle rotation, and performing YOLOv12-X detection on each transformed image. All detections are mapped back to the original image coordinate system by inverse transformation and fused with the original detection frames to obtain a more robust bounding box, which is used as a segmentation cue input to SAM2.

During the fusion process, they use the Hungarian algorithm to match the detected frames under different transformations by the intersection and concurrency ratio, and only retain the frames with  $\text{IoU} > 0.3$  and with area changes within a reasonable range. All matched frames for each target are weighted and averaged according to the confidence level to generate the final bounding box. The fusion results are fed into the SAM2 to obtain a more accurate instance segmentation mask. This method improves detection stability and helps improve the performance of metrics, such as HOTA, MOTA, and IDF1, in multi-target segmentation tasks.

### 5.2.3. Results

The team conducted a series of experiments on the MouseSIS dataset to evaluate the impact of different combinations of detection and segmentation models on the performance of multi-target tracking and segmentation (MOTS).

Table 2 shows the impact of different detector and segmentation model combinations on MOTS performance metrics. Replacing the segmentation module from the original SAM to the SAM2 under the YOLOv8 detector yields substantial improvements: HOTA increases by 8.44%, MOTA by 10.74%, and IDF1 by 11.98%. This indicates that the enhanced segmentation model significantly improves the quality of instance masks, which in turn benefits tracking accuracy. Building on this, replacing the detector from YOLOv8 to YOLOv12 results in additional gains of 0.69% in HOTA, 1.69% in MOTA, and 1.72% in IDF1, highlighting the importance of more precise and stable detection boxes in supporting segmentation and target association.

Furthermore, the use of CLAHE-based contrast enhancement, indicated by an asterisk in Tab. 2 as YOLOv12\*, results in additional improvements of 0.77% in HOTA, 0.55% in MOTA, and 1.37% in IDF1. These results demonstrate that progressive enhancements of the segmentation model, detection accuracy, and input quality collectively contribute to better MOTS performance.

Based on the use of the YOLOv12 detector, the SAM2 segmentation model, and CLAHE preprocessing, the team further explored the effect of replacing the original XMem tracker with XMem-no-sensory weights (Tab. 3). To ensure

Det. Model	Seg. Model	Tracker	HOTA↑	MOTA↑	IDF1↑
Yolov8	SAM	XMem	0.431	0.445	0.500
Yolov8	SAM2	XMem	0.516	0.552	0.620
Yolov12	SAM2	XMem	0.522	0.569	0.637
Yolov12*	SAM2	XMem	0.530	0.575	0.651

Table 2. (Team 2) Comparison of MOTS performance of different model combinations on MouseSIS test dataset.

Det. Model	Seg. Model	HOTA↑	MOTA↑	IDF1↑
YOLOv12	SAM2	0.535	0.575	0.650
YOLOv12+Val	SAM2	0.540	0.572	0.675
YOLOv12+Val	SAM2 (fine-tuned)	0.542	0.669	0.695
YOLOv12+Val+TTA	SAM2 (fine-tuned)	0.569	0.688	0.748

Table 3. (Team 2) Impact of Training and Inference Strategies on MOTS Performance.

the adequacy of training, the configuration labeled “+val” in the table indicates that only the data after the 400th frame in the 33rd sequence is used as the validation set, and the rest of the original validation dataset data are used to train YOLOv12. This segmentation strategy can better utilize the labeled data and effectively improve model performance.

In addition, the team fine-tuned the SAM2 segmentation model on the MouseSIS dataset to make it more adaptive to the specific task scenarios (Tab. 3), resulting in an increase of MOTA by 9.65%, and the introduction of the test-time-enhancement (TTA) strategy for the detection frames continued to optimize the performance of the model, and the HOTA increased to 0.569, the MOTA reached 0.688, and the IDF1 reached 0.748. The experiment further verifies the effectiveness of the multi-stage optimization strategy in the multi-target segmentation tracking task.

### 5.3. Team 3: mysterypeople

#### 5.3.1. Introduction

The team proposes a hybrid approach combining the event-image fusion segmentation framework EvInsMOS [22] and the tracking-by-detection methodology inspired by the XMem-based pipeline introduced in the MouseSIS benchmark [8].

#### 5.3.2. Method Overview

As shown in Fig. 3, the pipeline integrates the strengths of both segmentation and tracking:

- The team employs the EvInsMOS model as the backbone segmentation network, which fuses texture features from grayscale images and motion cues from event voxels.
- To reduce missed instance detection and improve spatial localization, they enhance the decoder with an additional bounding box regression head supervised by  $L^1$  loss.
- For temporal consistency and identity assignment, they apply an XMem-based tracker [4] to the segmentation outputs, enabling cross-frame association.

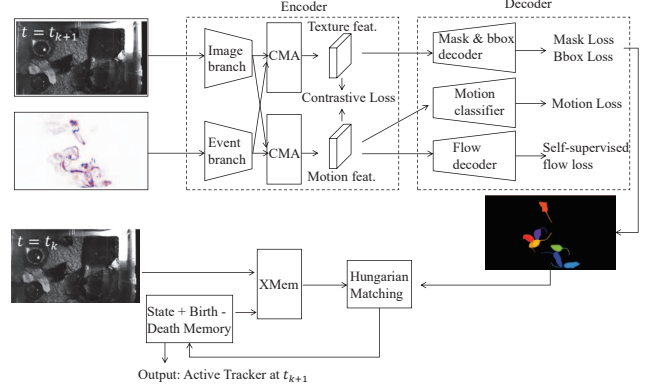


Figure 3. Team 3. Overall pipeline of the method combining EvInsMOS and XMem tracking.

#### 5.3.3. Segmentation Network Design

**Encoder.** The encoder consists of two modality-specific branches:

- **Image Branch:** A ResNet-based encoder extracts high-resolution texture features  $f_I$ .
- **Event Branch:** Event streams are voxelized into  $H \times W \times B$  tensors (with  $B = 10$  bins), following Eq. (1) from [22]. These are processed by a lightweight CNN to obtain motion features  $f_E$ .

These two feature maps are augmented via the **Cross-Modal Masked Attention (CMA)** module:

$$f_T = (f_E \odot f_{EM}) \cdot \text{softmax}(f_I^\top (f_E \odot f_{EM}) / \tau) + f_I \quad (1)$$

$$f_M = f_I \cdot \text{softmax}((f_E \odot f_{EM})^\top f_I / \tau) + f_E \quad (2)$$

where  $f_{EM}$  is an event mask,  $\tau$  is a learnable temperature, and  $\odot$  denotes element-wise multiplication.

**Contrastive Learning Objective:** The team adopts a multi-frame InfoNCE-based contrastive learning strategy, following the design in EvInsMOS [22]. Specifically:

- **Positive pairs** are constructed between feature representations of the same modality (either  $f_T$  or  $f_M$ ) across adjacent frames in a batch.
- **Negative pairs** include features from different modalities (i.e., motion vs. texture) as well as from non-adjacent samples across the batch.

The goal is to enforce inter-frame temporal consistency for each modality while encouraging decorrelation between texture and motion modalities. The total contrastive loss is:

$$\mathcal{L}_{cl} = -\frac{1}{B} \sum_{b=1}^B \log \left( \frac{FC_T^b + FC_M^b}{SS_T^b + SS_M^b + CS_{T,M}^b} \right) \quad (3)$$

In this formulation,  $B$  denotes the number of samples in the mini-batch. The terms are defined as follows:

- $FC_T^b$  and  $FC_M^b$  represent the feature consistency scores between the current and reference frame within the texture and motion modalities, respectively. These scores correspond to positive pairs.
  - $SS_T^b$  and  $SS_M^b$  denote the self-similarity scores within the same modality but from unrelated samples, used as intra-modality negative pairs.
  - $CS_{T,M}^b$  is the cross-modality similarity score between texture and motion features across the batch, used to penalize shared representations between distinct modalities.
- This formulation ensures temporal coherence within modalities while enforcing modality-specific representations.

**Decoder.** The decoder adopts a query-based design inspired by Mask2Former [3], which decouples segmentation and classification into separate parallel branches. The team generates  $n$  fixed learnable queries that attend to different object instances.

- **Mask Decoder:** Takes the augmented texture feature  $f_T$  and projects each query embedding into a segmentation mask  $\hat{S}_i$ . Here,  $n$  denotes the total number of queries.
- **Motion Classifier:** Takes the augmented motion feature  $f_M$  and predicts a binary motion score  $\mathbf{m}_i \in \{0, 1\}$  for each query, indicating whether the corresponding instance is moving.

Decoupling these tasks is beneficial because motion state may not perfectly align with spatial contours, especially in the presence of camera-induced parallax. This separation allows motion classification to benefit from motion-specific cues and segmentation to focus on spatial accuracy.

Methods	HOTA↑	MOTA↑	IDF1↑
ModelMixSort (baseline) [8]	0.43	0.45	0.50
EvInsMOS + bbox + Hungarian	0.509	0.498	0.605
EvInsMOS + XMem tracker	0.524	0.579	0.640
EvInsMOS + bbox + XMem ( <i>Team 3</i> )	<b>0.542</b>	<b>0.594</b>	<b>0.669</b>

Table 4. (*Team 3*) Performance comparison on original resolution (1280×720 px) MouseSIS test set.

**Training:** For each frame, ground truth instance masks  $S_{gt}^{(i)}$  and their motion labels  $c_{gt}^{(i)}$  are assigned to predicted queries using Hungarian matching  $\rho(i)$  based on spatial IoU. The combined loss is:

$$\mathcal{L}_{\text{mos}} = \sum_{i=1}^n \left( \mathcal{L}_{\text{ce}}(\mathbf{m}_{\rho(i)}, c_{gt}^{(i)}) + \mathbb{1}_{c_{gt}^{(i)}=1} \cdot \mathcal{L}_{\text{mask}}(\hat{S}_{\rho(i)}, S_{gt}^{(i)}) \right) \quad (4)$$

where  $\mathcal{L}_{\text{mask}}$  is a mixture of focal and dice losses, and  $\mathcal{L}_{\text{ce}}$  is binary cross entropy for motion label prediction.

Also, the team adds a bounding box regression loss:

$$\mathcal{L}_{\text{bbox}} = \sum_{i=1}^n \left\| \hat{b}_i - b_i^{gt} \right\|_1 \quad (5)$$

**Inference:** At test time, all  $n$  predicted masks  $\hat{S}_i$  and motion scores  $\mathbf{m}_i$  are computed. Masks with  $\text{softmax}(\mathbf{m}_i) > \theta$  are retained as the final  $m$  moving instance predictions. This enables the model to adaptively determine the number of foreground instances.

To further enhance the decoder’s sensitivity to motion boundaries, the team incorporates an optical flow-guided feature modulation mechanism. An unsupervised flow estimator computes the optical flow between adjacent grayscale frames. The resulting flow field is used to warp and align decoder-level feature maps across time. This flow-guided alignment enhances temporal coherence and sharpens motion contours, particularly in occlusion-prone or fast-motion regions. The warped feature  $f_{\text{warp}}$  is fused with the decoder output via attention-based gating, refining the segmentation quality without requiring additional supervision.

### 5.3.4. Tracking and ID Association

The team uses the XMem [4] tracker, specifically they:

1. Generate per-frame masks from EvInsMOS.
2. Feed masks and frames to XMem for propagation.
3. Use Hungarian matching based on IoU to align new masks with existing memory trackers.

### 5.3.5. Implementation Details

The implementation is in PyTorch 2.5.1, on four NVIDIA A40 GPUs with batch size 8, Adam optimizer, an initial learning rate of  $1 \times 10^{-4}$  trained for 300,000 iterations. For the XMem-based tracker, the team uses the following hyperparameter configuration: the maximum age of a tracker (‘max\_age’) is set to 1, ‘min\_hits = 3’. The IoU threshold is set to 0.5 (‘iou\_threshold = 0.5’).

### 5.3.6. Results

The team compares their method against three alternative settings: (1) ModelMixSort baseline as proposed in [8]; (2) EvInsMOS with added bounding box regression and Hungarian matching for association; (3) EvInsMOS combined with XMem-based tracker. Their full model integrates both the enhanced decoder and XMem tracking.

From Table 4, it can be observed that the method outperforms all baselines across all three metrics. The addition of the bounding box regression head (comparing row 3 with row 4) contributes to better localization and motion estimation. Replacing Hungarian matching with the memory-based XMem tracker (comparing row 2 with row 4) boosts identity preservation as reflected in IDF1. Combining both enhancements results in the best performance.

## Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy – EXC 2002/1 “Science of Intelligence” – project number 390523135.

## References

- [1] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear MOT metrics. *EURASIP J. Image and Video Processing*, pages 1–10, 2008. Article ID: 246309.
- [2] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *IEEE Int. Conf. Image Process. (ICIP)*, pages 3464–3468, 2016.
- [3] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 1290–1299, 2022.
- [4] Ho Kei Cheng and Alexander G Schwing. XMem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 640–658, 2022.
- [5] Thomas Finateu, Atsumi Niwa, Daniel Matolin, Koya Tsuchimoto, Andrea Mascheroni, Etienne Reynaud, Poo-ria Mostafalu, Frederick Brady, Ludovic Chotard, Florian LeGoff, Hirotsugu Takahashi, Hayato Wakabayashi, Yusuke Oike, and Christoph Posch. A 1280x720 back-illuminated stacked temporal contrast event-based vision sensor with 4.86 $\mu$ m pixels, 1.066Geps readout, programmable event-rate controller and compressive data-formatting pipeline. In *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, pages 112–114, 2020.
- [6] Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew Davison, Jörg Conrad, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(1):154–180, 2022.
- [7] Rafael C. Gonzalez and Richard Eugene Woods. *Digital Image Processing*. Pearson Education, 2009.
- [8] Friedhelm Hamann, Hanxiong Li, Paul Mieske, Lars Lewejohann, and Guillermo Gallego. MouseSIS: A frames-and-events dataset for space-time instance segmentation of mice. In *Eur. Conf. Comput. Vis. Workshops (ECCVW)*, pages 156–173, 2024.
- [9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Int. Conf. Comput. Vis. (ICCV)*, pages 1026–1034, 2015.
- [10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. *Int. Conf. Learn. Representations (ICLR)*, 1(2):3, 2022.
- [11] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023.
- [12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 3992–4003. IEEE, 2023.
- [13] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128 $\times$ 128 120 dB 15  $\mu$ s latency asynchronous temporal contrast vision sensor. *IEEE J. Solid-State Circuits*, 43(2):566–576, 2008.
- [14] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. HOTA: A higher order metric for evaluating multi-object tracking. *Int. J. Comput. Vis.*, 129(2):548–578, 2021.
- [15] Keiron O’Shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.
- [16] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloé Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross B. Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.
- [17] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(6):1964–1980, 2021.
- [18] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 17–35, 2016.
- [19] Jenny Seidenschwarz, Guillem Brasó, Victor Castro Serrano, Ismail Elezi, and Laura Leal-Taixé. Simple cues lead to a strong multi-object tracker. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 13813–13823, 2023.
- [20] Yunjie Tian, Qixiang Ye, and David S. Doermann. Yolov12: Attention-centric real-time object detectors. *CoRR*, abs/2502.12524, 2025.
- [21] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. MOTs: Multi-object tracking and segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 7942–7951, 2019.
- [22] Zhexiong Wan, Bin Fan, Le Hui, Yuchao Dai, and Gim Hee Lee. Instance-level moving object segmentation from a single image with events. *Int. J. Comput. Vis.*, 2025.
- [23] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Int. Conf. Comput. Vis. (ICCV)*, pages 5187–5196, 2019.
- [24] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 492–510, 2022.
- [25] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-object tracking by associating every detection box. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 1–21, 2022.