

SAMPLE: Semantic Alignment through Temporal-Adaptive Multimodal Prompt Learning for Event-Based Open-Vocabulary Action Recognition

Supplementary Material

We provide supplementary material to offer additional details and further qualitative analyses that complement the main paper. The contents are organized as follows:

- Dataset Details
- Evaluation metrics
- Efficient Analysis
- Additional Ablation Studies
- Additional Qualitative Results

1. Dataset Details

Our analysis centers on four well-established benchmarks commonly used for action recognition in event-based data: HARDVS [6], DVS128Gesture [2], SeAct [7] and PAF [3]. **HARDVS** The HARDVS dataset serves as the first large-scale benchmark tailored specifically for human activity recognition (HAR) using event cameras. Captured with a DAVIS346 camera at a resolution of 346×260 pixels, it comprises 107,646 unique event sequences covering 300 diverse activity categories, including "drinking water", "brushing teeth," and "writing." Each sequence spans an average duration of 5–10 seconds, culminating in a dataset of approximately 1,076 hours of recorded activity.

DVS128Gesture The DVS128Gesture dataset consists of 1,342 instances of 11 distinct hand and arm gesture categories (e.g., hand waving, arm rotations, and air drum), performed by 29 unique subjects under three lighting conditions: natural sunlight, fluorescent lighting, and LED lighting. These recordings were made with the DVS128 camera, which offers a resolution of 128×128 pixels and generates asynchronous events with a high dynamic range of up to 120 dB. Each gesture sequence lasts approximately 6–10 seconds.

SeAct The SeAct dataset is a semantically rich benchmark designed for event-text action recognition, featuring detailed caption-level annotations for each action. The dataset is collected using a DAVIS346 event camera with a resolution of 346×260 pixels and includes 58 distinct actions grouped into four thematic categories, as illustrated in the accompanying images.

PAF The PAF (Pedestrian Action and Fall) dataset focuses on safety-critical applications such as fall detection and pedestrian activity recognition. This dataset, recorded with the DAVIS346redColor sensor, includes 450 event recordings across 10 distinct action categories: "arm-crossing," "jumping," "kicking," "getting-up," "throwing," "walking," and more. Each recording lasts approximately 5 seconds, providing a total of 3,150 seconds (52.5 minutes) of annotated event data.

tated event data.

2. Evaluation metrics

In this section, we show evaluation metrics of SAMPLE for fully-supervised setting, few-shot, base-to-novel, and zero-shot experimental settings in the decreasing supervision level. The event data is processed using the Adaptive Fine-grained Event (AFE) representation [7], which converts the raw events into three-channel event frames. These frames are further pre-processed to a spatial resolution of 224×224 . For this setting, we utilize 8 sampled event frames for model input.

(A) Fully-Supervised Fully-supervised setting aims to assess our method's task-specific performance, SAMPLE is trained on the HARDVS, DVS128Gesture, SeAct and PAF datasets and evaluated on their entire validation set respectively.

(B) Few-Shot We further evaluate SAMPLE's generalization capability within a few-shot learning framework, assessing its performance with a limited number of labeled samples per class. The few-shot setting involves generating a generalized K-shot split, where each class is represented by K samples. We specifically evaluate $K = 2, 4, 8$ and 16 shots across four datasets: HARDVS, DVS128Gesture, SeAct, and PAF. The models are tested on their respective full validation sets to assess performance.

(B) Base-to-Novel To evaluate SAMPLE's generalization capabilities for novel classes, we conduct experiments in a base-to-novel setting, following the approach outlined in [4]. We first split the dataset with the manner that the most frequently occurring categories are designated as base classes, while the less frequent categories serve as novel classes. And then our model is initially trained in a few-shot ($K = 16$) manner on base classes and then evaluated on both base and novel classes.

(B) Zero-shot In the zero-shot setting, models trained on the training set of the HARDVS dataset are evaluated on the evaluation sets of three distinct cross-datasets: DVS128Gesture, SeAct, and PAF, to comprehensively assess their generalization capability to unseen datasets.

3. Efficient Analysis

We compare the computational efficiency of SAMPLE with other methods in Table 1. Recent CLIP-based RGB video recognition methods, such as EZ-CLIP [1] and M2-CLIP [5], employ parameter-efficient fine-tuning techniques to

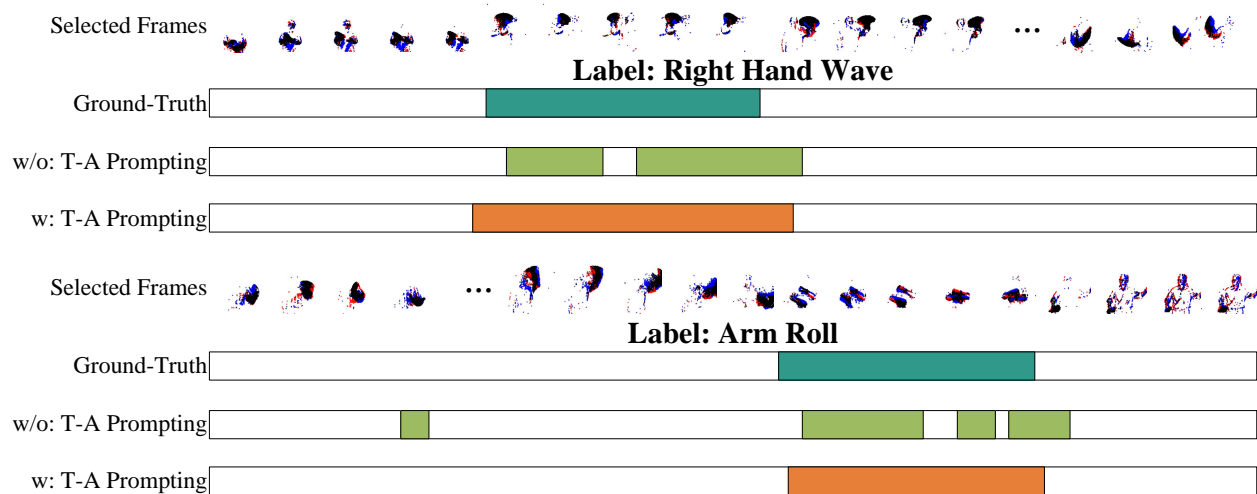


Figure 1. Visualization of temporal action localization results for two gesture categories, "Right Hand Wave" and "Arm Roll," from the DVS128Gesture dataset. The ground truth annotations indicate the precise start and end times of the gestures. Without T-A Prompting, the predicted segments are dispersed and include false positives beyond the gesture duration. In contrast, with T-A Prompting, the predicted segments align closely with the ground truth. Selected frames illustrate the visual content corresponding to the annotated intervals.

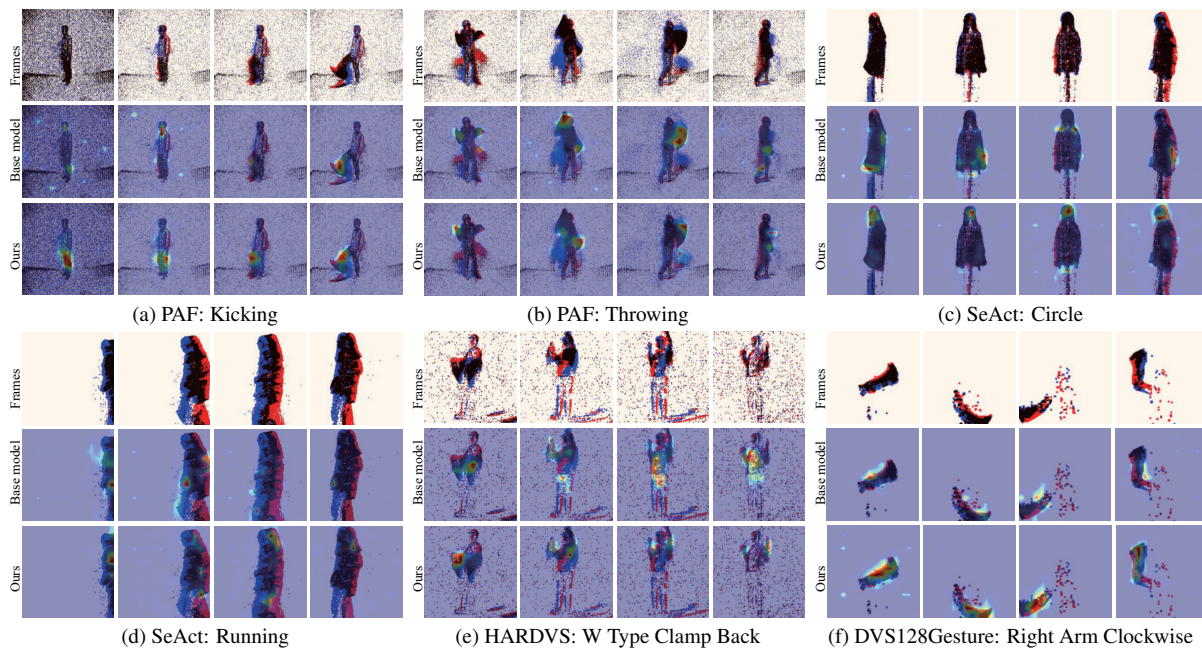


Figure 2. Qualitative comparison of attention maps for different actions between base model and our proposed SAMPLE across PAF, SeAct, HARDVS and DVS128Gesture datasets.

091 adapt the CLIP model but lack specialized adaptations tai-
 092 lored to event data. In contrast, ExACT [7] is the first work
 093 to leverage language information to assist event-based ac-
 094 tion recognition by combining a pre-trained CLIP text en-
 095 coder with a fully trainable, tailored event encoder.

096 SAMPLE introduces slightly more tunable parameters
 097 than EZ-CLIP (8.681 million vs. 5.200 million) due to
 098 its specific design tailored for event data. However, it re-

mains significantly more parameter-efficient compared to
 the previous state-of-the-art method, ExACT, which re-
 quires 47.917 million tunable parameters due to partial fine-
 tuning. This substantial reduction in parameters not only
 enhances computational throughput but also ensures com-
 parable computational complexity, demonstrating the effi-
 ciency of SAMPLE.

099
 100
 101
 102
 103
 104
 105

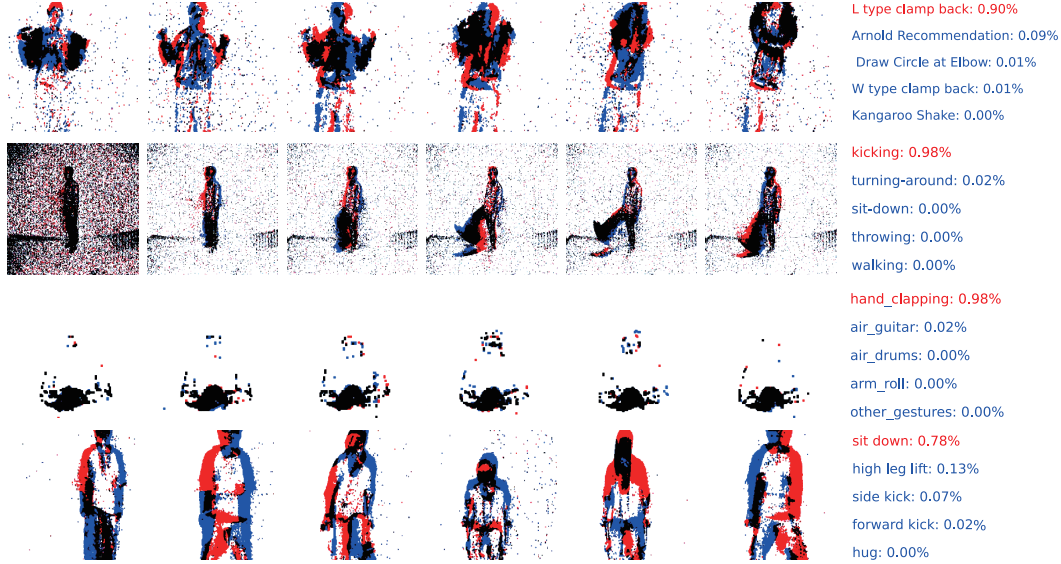


Figure 3. Visualization of the top-5 predicted results on the HARDVS, PAF, DVS128Gesture and SeAct datasets (from top to bottom).

Method	TP \uparrow	Total Params (M) \downarrow	Tunable Params (M) \downarrow
ExACT	78.97	283.732	47.917
M2-CLIP	59.54	421.000	16.000
EZ-CLIP	288.79	88.400	5.200
Ours	<u>233.40</u>	<u>133.090</u>	<u>8.681</u>

Table 1. Compute comparison of SAMPLE. Throughput per view (TP) is measured using a single RTX4090D GPU.

4. Additional Ablation Studies

Evaluating Temporal Sensitivity in Temporal-Adaptive Prompting We evaluate the temporal sensitivity of our Temporal-Adaptive Prompting (T-A Prompting) mechanism on the temporal action detection (TAD) task, which focuses on identifying actions and their precise temporal boundaries in untrimmed sequences. Using the DVS128Gesture dataset with ground-truth annotations of gesture labels and their start and stop times, we compare the model’s performance with and without T-A Prompting to measure its effectiveness in accurate action localization over time.

The results are summarized in Table 2, where the average mAP improves by approximately 18.95% when T-A Prompting is incorporated, demonstrating a significant enhancement across all IoU thresholds. This indicates that T-A Prompting substantially boosts the model’s ability to capture temporal dynamics effectively.

To further demonstrate the impact of T-A Prompting, we provide a visualization of action localization results for two gesture categories from the DVS128Gesture dataset: ‘right Hand Wave’ and ‘Arm Roll’. Without T-A Prompting, the predicted segments are dispersed and frequently include false positives beyond the actual gesture

duration. In contrast, incorporating T-A Prompting results in predictions that closely align with the ground truth, precisely identifying the start and end times of each gesture. T-A Prompting ensures temporal alignment, reduces noise in predictions and improves overall localization accuracy. This underscores the necessity of temporal adaptivity in action recognition tasks.

Method	IOU				Avg mAP(%)
	0.1	0.3	0.5	0.7	
w/o T-A prompting	0.55	0.46	0.28	0.19	0.37
w T-A prompting	0.72	0.58	0.39	0.27	0.49

Table 2. Comparison of mAP values at different IOU thresholds of [0.1,0.3,0.5,0.7] on DVS128Gesture dataset with and without Temporal-Adaptive (T-A) prompting.

Effectiveness of Prompt Depth We investigate the impact of prompt depth on performance. Figure 4 illustrates that as the prompt depth increases, the model’s performance improves, reaching optimal accuracy at a depth of 12. This suggests that deeper prompts allow the model to capture more complex hierarchical features, enhancing its ability to represent and recognize actions in event data.

5. Additional Qualitative Results

Attention Map Visualizations In Fig. 2, we provide an additional visualization of attention maps for the HARDVS, PAF, DVS128Gesture, and SeAct datasets. Compared to the base model, our method demonstrates a more precise focus on the critical areas where the action occurs. For instance, in the ‘Kicking’ action from the DVS128Gesture dataset, our method accurately highlights the body part (leg) directly involved in the action. Similarly, for the ‘Throwing’ action in the PAF dataset, the attention is more effectively

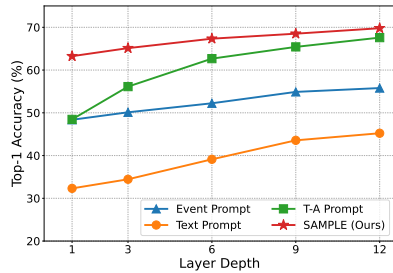


Figure 4. Comparison of Top-1 accuracy (%) across varying layer depths for different prompting strategies: Event Prompt only, Text Prompt, T-A Prompt, and the proposed SAMPLE framework.

concentrated on the object and regions associated with the throwing motion. This enhanced focus on action-relevant regions underscores the model’s ability to capture fine-grained spatiotemporal details, significantly improving its interpretability and performance in action recognition tasks.

Top-5 Predicted Results In Fig. 3, we present a visualization of the top-5 predicted results for the HARDVS, PAF, DVS128Gesture, and SeAct datasets, accompanied by selected event frames. These predictions are represented by probability scores for the top-5 categories in each dataset, with the scores summing to 1.

The figure shows that the top-2 predictions in the top-5 generally exceeds 0.95, demonstrating the model’s ability to capture target semantics and offer reasonable predictions among plausible options.. And the semantic proximity of the top-5 categories, such as ‘L type clamp back’ and ‘W type clamp back’ in the HARDVS dataset, reflects the model’s ability to map visual and textual features into a shared representation space. This semantic alignment is particularly critical in zero-shot learning scenarios or tasks with blurred inter-class boundaries

References

- [1] Shahzad Ahmad, Sukalpa Chanda, and Yogesh S Rawat. Ez-clip: Efficient zeroshot video action recognition. *arXiv*, 2023. 1
- [2] Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, et al. A low power, fully event-based gesture recognition system. In *CVPR*, pages 7243–7252, 2017. 1
- [3] Shu Miao, Guang Chen, Xiangyu Ning, Yang Zi, Kejia Ren, Zhenshan Bing, and Alois Knoll. Neuromorphic vision datasets for pedestrian detection, action recognition, and fall detection. *Frontiers in Neurorobotics*, 13:38, 2019. 1
- [4] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *CVPR*, 2023. 1
- [5] Mengmeng Wang, Jiazheng Xing, Boyuan Jiang, Jun Chen, Jianbiao Mei, Xingxing Zuo, Guang Dai, Jingdong Wang, and Yong Liu. M2-clip: A multimodal, multi-task adapting framework for video action recognition. *arXiv*, 2024. 1
- [6] Xiao Wang, Zongzhen Wu, Bo Jiang, Zhimin Bao, Lin Zhu, Guoqi Li, Yaowei Wang, and Yonghong Tian. Hardvs: Revisiting human activity recognition with dynamic vision sensors. *arXiv*, 2022. 1
- [7] Jiazhou Zhou, Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. Exact: Language-guided conceptual reasoning and uncertainty estimation for event-based action recognition and more. In *CVPR*, pages 18633–18643, 2024. 1, 2