E-CLIP: Towards Label-efficient Event-based Open-world Understanding by CLIP –Supplementary Material–

Jiazhou Zhou*, Xu Zheng*, *Student Member, IEEE*, Yuanhuiyi Lyu, Lin Wang†, *Member, IEEE*
Project homepage: https://vlislab22.github.io/ECLIP/

APPENDEX

1 IMPLEMENTATION DETAILS

1.1 Event Representation

Event cameras capture object movement by detecting temporal independence and recording pixel-level brightness changes. The event stream, denoted as $\mathcal{E} = e_i(x_i, y_i, t_i, p_i)$, reflects the brightness change e_i of a pixel at the timestamp t_i, with coordinates (x_i, y_i), and polarity $p_i \in 1, -1$. Here, 1 and -1 represent the positive and negative intensity change of brightness, respectively.

To obtain a grid-like tensor as the input, we convert the event stream into a sequence of frames. First, we normalize the length of the event stream $\mathcal{E}$ to a fixed amount P using zero-padding or taking the first P events. Then, we group every Q consecutive event data in the normalized event stream $\mathcal{E}^{'}$ to obtain event parts $\mathcal{E}^{''} \in \mathbb{R}^{T\times 4\times P}$, where $T = P/Q$ and denotes the number of event frames. We then transform these event parts into histograms $h \in \mathbb{R}^{T\times H\times W\times 2}$ by counting the amount of positive and negative events per pixel, where H and W represent the height and width of an event frame, respectively. Finally, to attain the grid-like 3-channel input $I_e \in \mathbb{R}^{T\times H\times W\times 3}$, we colorize the tth event histogram h_t by multiplying the predefined red color map $[0, 255, 255]$ and blue color map $[255, 255, 0]$ with the positive and negative event histogram respectively and merge them by pixel-wise addition, which can be formulated as follow:

$$I_{e,t} = h_t^p[0, 255, 255]^T + h_t^n[255, 255, 0]^T, \quad (1)$$

where the $I_{e,t}$ represents the event input for the tth event frame and $t = 0, 1, ..., T$; h_t^p and h_t^n denotes the positive and negative event histogram respectively. Finally, the generated event frame input is resized into the 224×224 resolution for adapting to the ViT setting, generating the final event input $I_e^{'} \in \mathbb{R}^{T\times 224\times 224\times 3}$.

The above grid-like event representation is simple yet effective since the generated event image resembles the edge image, whose data distribution is much closer to the pre-trained natural images so that the transfer pressure is erased and the modality gap is bridged.

- *J. Zhou, X. Zheng, Y. Lyu are with the AI Thrust, HKUST(GZ), Guangzhou, China. E-mail: zhengxu128@gmail.com, {jiazhouzhou, yuanhuiyilv}@kust-gz.edu.cn*
- *L. Wang is with the AI Thrust, HKUST(GZ), Guangzhou, and Dept. of Computer Science and Engineering, HKUST, Hong Kong SAR, China. E-mail: linwang@ust.hk*

Manuscript received April 19, 2022; revised August 26, 2022.
*(*Equal contribution, †Corresponding author: Lin Wang)*

Length of text learnable prompt	Top1 Accuracy	
	N-Caltech101	N-MNIST
4	92.50	97.20
8	92.68	98.50
16	92.92	98.71

TABLE 1
Ablation study on the length of text learnable prompt.

Length of event modality prompt	Top1 Accuracy	
	N-Caltech101	N-MNIST
4	92.50	97.75
8	92.68	98.73
16	92.92	98.82

TABLE 2
Ablation study on the length of event modality prompts

2 ADDITIONAL EXPERIMENT RESULT

2.1 Accuracy Curve

Two significant ablation experiments conducted on the N-Caltech101 dataset of E-CLIP are chosen for analysis. These investigations include an ablation study that explores various combinations of the proposed module, as well as an ablation study that investigates the Hierarchical Triple Contrastive Alignment (HTCA) module. We present the loss curves of them in Fig. 1 to observe their accuracy change with varying ablation settings. As shown in Fig. 1, the accuracy of the model exhibits fluctuations that tend to rise as the number of training epochs increases. The observed discrepancy in the ultimate accuracy across several ablation versions of the model provides more evidence supporting the efficacy and logicality of the proposed model.

2.2 Visulization of Few-shot Event Recognition Results

Fig. 2 displays the accuracy curve with different evaluation settings, including zero-shot, few-shot, and fine-tuning on the N-Caltech101, NMIST, and N-Imagenet datasets respectively. We conduct a comparative analysis between E-CLIP and the existing EventCLIP on the N-Caltech101 and N-ImageNet datasets. Our findings indicate that E-CLIP outperforms its competition when provided with a minimum of five training examples. These results highlight the efficacy and label efficiency of the E-CLIP model.

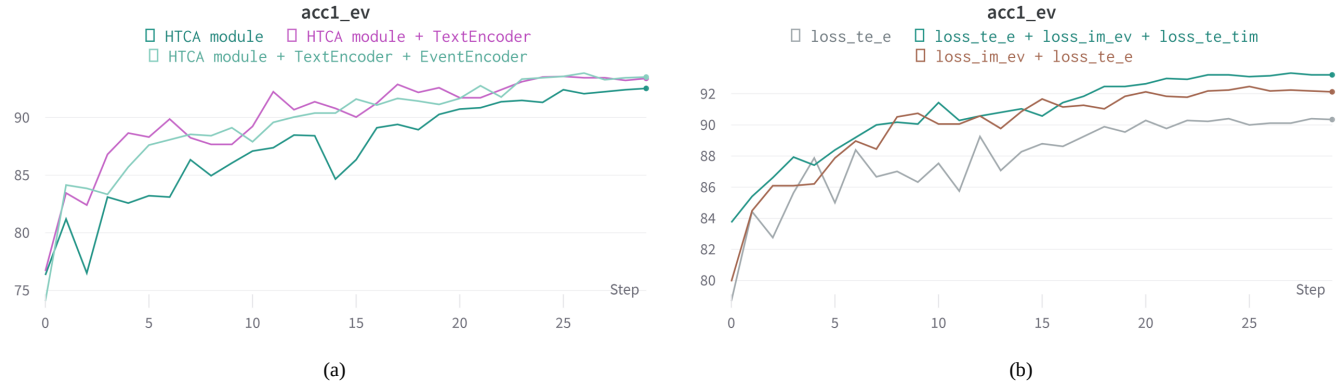

Fig. 1. Visulization of accuracy curve on two selected important ablation studies on the N-Caltech101 dataset of our proposed E-CLIP. (a)The ablation study on different combinations of the proposed module; (b)The ablation study on the HTCA module.

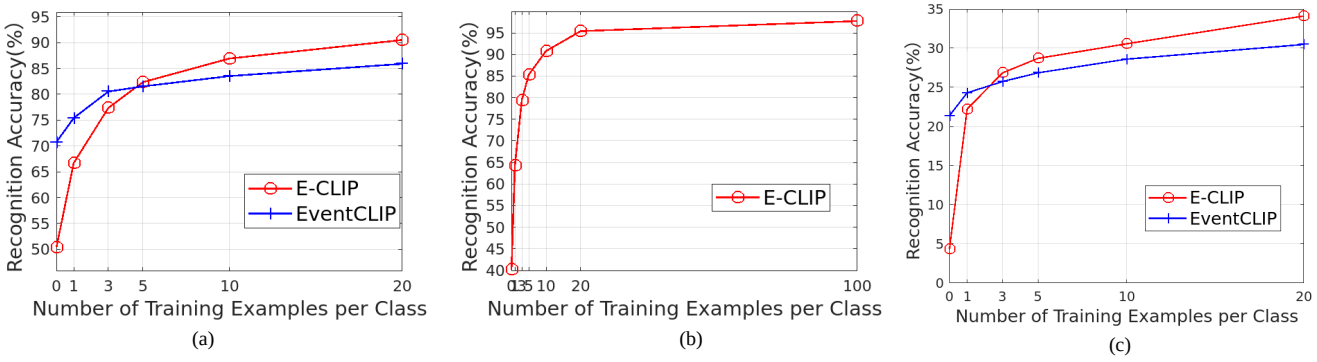

Fig. 2. Visulization of event recognition results with zero-shot, few-shot settings. (a) Results on the N-Caltech101 dataset; (b) Results on the NMIST dataset; (c)Results on the N-Imagenet dataset.

Setting	Text2Event			Imgae2Event		
	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10
0-shot	78.22	88.12	91.09	79.40	90.34	93.58
1-shot	81.19	95.05	98.02	82.75	96.04	99.01
2-shot	90.10	99.01	100.00	88.31	98.02	99.01
5-shot	97.03	100.00	100.00	90.97	99.01	100.00
10-shot	98.02	100.00	100.00	94.73	99.01	100.00
20-shot	99.01	100.00	100.00	96.70	100.00	100.00
Fine-tune	100.00	100.00	100.00	96.88	99.60	99.95

TABLE 3
Event Retrieval results on N-Caltech101 dataset with zero-shot, few-shot, and fine-tuning settings.

2.3 Full Numerical Results

In the main paper, we plot E-CLIP's ablation study on the length of the text learnable prompts, ablation study on the length of the event modality prompts, and Event retrieval results on N-Caltech101 dataset with zero-shot, few-shot, and fine-tuning settings in Fig. 7(a), Fig. 7(b) and Fig. 9 respectively. To ease future comparison, we report all those numbers in Tab. 1, Tab. 2, and Tab. 3.