EventBind: Learning a Unified Representation to Bind Them All for Event-based Open-world Understanding

Jiazhou Zhou[1], Xu Zheng[1], Yuanhuiyi Lyu[1], and Lin Wang[1,2]⋆

[1] Hong Kong University of Science and Technology, Guangzhou, China
{jiazhouzhou,yuanhuiyilv}@hkust-gz.edu.cn, zhengxu128@gmail.com
[2] Hong Kong University of Science and Technology, Hong Kong, China
linwang@ust.hk
Project Page: https://vlislab22.github.io/EventBind/

Appendix

1 Preliminary of CLIP

The Contrastive Language-Image Pre-training (CLIP) [13] model is a prominent vision-language model that combines an image encoder (ViT [4] or ResNet [6]), with a text encoder based on the Transformer [15] architecture. CLIP focuses on generating visual and text embeddings and employs a contrastive loss to align these embeddings in a unified feature space. Notably, CLIP's exceptional transferability is attributed to its pretraining on a large-scale dataset of more than four million image-text pairs.

For CLIP-based recognition, it usually employs a simple yet efficient method that takes N hand-crafted prompts like 'a photo of a $[cls]$' as the input for the text encoder to obtain the D-dimensional textual embedding $f^t \in \mathbb{R}^{N\times D}$, where $[cls]$ is the class name and N is the number of classes of the downstream dataset. Meanwhile, for a given image, the visual embedding $f^i \in \mathbb{R}^{1\times D}$ is obtained by the visual encoder. Then, the recognition logits $p \in \mathbb{R}^N$ can be obtained by calculating the similarity of text and visual embeddings by multiplication:

$$p = \mathrm{Softmax}(f^i(f^t)^T), \tag{1}$$

where Softmax(.) denotes the Softmax function and p represents the predicted probability of the N classes. Finally, the highest scores of the logits p is regarded as the final prediction P:

$$P = \arg\max_{i\in\mathcal{N}} p_i \tag{2}$$

⋆ Corresponding author

2 Implementation Details

2.1 Event Frame-like 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.

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:

$$I_{e,t} = h_t^p[0, 255, 255]^T + h_t^n[255, 255, 0]^T, \tag{3}$$

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.

2.2 Additional Dataset Settings

N-ImageNet [7] is derived from the ImageNet-1K dataset, where the RGB images are displayed on a monitor and captured by a moving event camera. It includes 1,781,167 event streams with 480 × 640 resolution across 1,000 unique object classes. **N-Caltech101** [5] contains event streams captured by an event camera in front of a mobile 180 × 240 ATIS system [12] with the LCD monitor presenting the original RGB images in Caltech101. There are 8,246 samples comprising 300 ms in length, covering 101 different types of items. **N-MNIST** is created by displaying a moving image from the MNIST dataset on the ATIS system with the LCD monitor. It contains 70,000 event data samples covering 10 handwritten numbers from 0 to 9.

Backbone	N-Imagnet	N-MNIST	N-Caltech101
ViT-B-16	300,000	300	150,000
ViT-B-32	300,000	400	150,000
ViT-L-14	400,000	1,000	200,000

Table 1: The aggregated event counts per frame N for different backbones on three datasets.

Ablation settings		Base		New	
Learnable	Hand-crafted	N-Caltech101	N-MNIST	N-Caltech101	N-MNIST
✗	✓	90.57	91.03	**72.55**	18.80
✓	✗	91.01	93.22	61.85	46.02
✓	✓	**91.23**	**93.56**	68.02	**49.92**

Table 2: Ablation study on hybrid text prompts module.

2.3 Additional Experimental Settings

We use "A drafted image of a $[CLS]$" as the hand-crafted text prompts template. The Pytorch [11] framework serves as the foundation for all experiments. The initial learning rates are set to 1e-5 for the N-Caltech101 dataset and 1e-6 for the N-Imagenet and N-MNIST datasets. The weight decay is 2e-4. CosineAnnealingLR [9] learning rate schedule is utilized, and the minimal learning rate is 1e-8. All few-shot and fine-tuning experiments are trained for 30 epochs with Adam [8] optimizer. Unless specified otherwise, the ablation study is conducted on the N-Caltech101/Caltech101 datasets utilizing ViT-B-16 [3] image encoder and the Transformer-based text encoder [15] as the backbone. The event encoder is initialized with the CLIP image encoder's pre-trained weights and is fine-tuned during training. We choose the aggregated event counts per frame N based on the best EventBind's zero-shot performance. Tab. 1 represents the value of N set for different backbones on three datasets.

3 Additional experiment result

The effectiveness of hybrid text prompts We ablated the hybrid text prompts to evaluate their impact on fine-tuning performance and generalization ability. The experiment follows the setting proposed in [18], where the entire dataset is divided into the base and new datasets, each containing half of the object classes. For each ablation setting, the model is fine-tuned on the base set without being exposed to half of the event classes in the new set. The evaluation is then performed on the new set to assess the generalization ability.

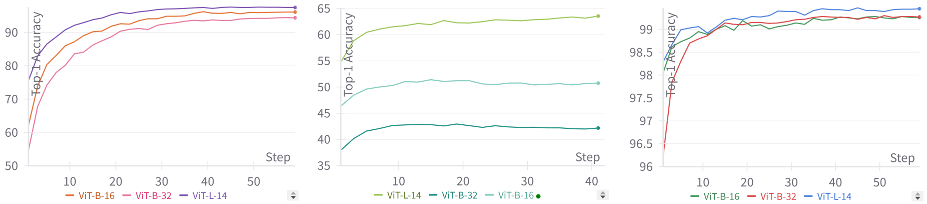
As shown in Tab. 2, the hand-crafted text prompts achieve an impressive 72.55% Top-1 accuracy on the new set for the N-Caltech101 dataset, showcasing its remarkable zero-shot ability. In contrast, while learnable text prompts boost fine-tuning performance on the base set, they lead to decreased accuracy on the

	EventBind		EventCLIP	
	N-Calthech101	N-MINIST	N-Calthech101	N-MINIST
0-shot	**61.80**	**56.81**	58.82	48.72
1-shot	**74.96**	**74.64**	70.53	74.62
2-shot	**79.78**	**82.89**	77.71	82.78

Table 3: Zero-shot, 1-shot & 2-shot results.

Model	E2VID [14]	SSL-E2VID [10]	Wang et al. [16]	Ev-LaFOR [2]	**EventBind(Ours)**
Top1 Accuracy	61.7	25.3	48.5	85.56	**86.02**

Table 4: Open-vocabulary results on N-Caltech101.



(a) N-Caltech101 (b) N-Imagenet (c) N-MNIST

Fig. 1: The Accuracy curves with three ViT backbones on the N-MINIST, N-Caltech101, and N-Imagenet datasets.

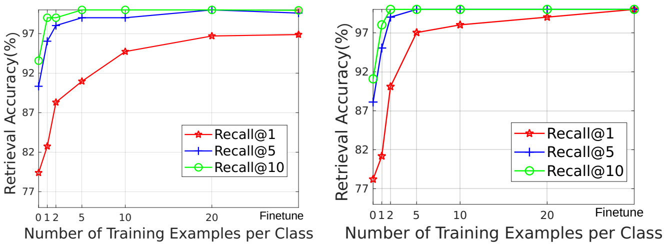
new set due to limited generalization capabilities. *In contrast, our proposed hybrid text prompts module, which integrates both learnable and manually-crafted text prompts, attains the highest fine-tuning accuracy of 91.23% while exhibiting a smaller decrease in zero-shot performance relative to solely utilizing single hand-crafted text prompts.* (**-10.70% V.S -4.53%** for N-Caltech101). For the N-MNIST dataset, it's expected that the model exhibits low accuracy on the base dataset with hand-crafted text prompts. This stems from the subpar image recognition capability of the original CLIP on N-MNIST, where zero-shot CLIP results are 10% inferior to those of the linear probe on ResNet50 for MNIST. [13]). *Our model, equipped with the hybrid text prompts module, secures top-one accuracy on both the base dataset (93.56%) and the new dataset (98.86%) for N-MNIST, underscoring the effectiveness of our proposed module.*

Zero-shot, 1-shot, 2-shot results comparing with EventCLIP. We provide results in Tab.3. Since the dataset split of EventCLIP is unavailable, we evaluated it on our dataset split with ViT-L-14 backbones. Our EventBind achieves better zero-shot, 1-shot, and 2-shot performance compared with EventCLIP.

Open-vocabulary Recognition Result For open-vocabulary event recognition, the text classes are arbitrary rather than restricted to the training classes. We follow the dataset split based on its open-source repository. The results are in Tab.4. Our EventBind achieves the SoTA performance of **86.02%**, proving its superiority for open-vocabulary recognition.

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



(a) Image-to-event retrieval.

(b) Text-to-event retrieval.

Fig. 2: Event retrieval results on N-Caltech101 dataset with zero-shot, few-shot, and fine-tuning settings.

Accuracy Curve We present the training accuracy curves trained by employing three ViT backbones on the N-MINIST, N-Caltech101, and N-Imagenet datasets in Fig. 1. As the number of training epochs increases, the accuracy of the model increases steadily, demonstrating the stability and reproducibility of EventBind.
Event Retrieval Numerical Results We utilize Recall@1, Recall@5, and Recall@10 metrics commonly employed in retrieval tasks [1, 17]. In Fig. 2, our EventBind shows remarkable retrieval performance (99.01% Recall@1 for text query and 96.70% Recall@1 for image query) with only 20 shot training examples, demonstrating its remarkable capabilities in few-shot learning. Notably, our model excels in text-to-event and image-to-event retrieval, achieving recall rates close to 100.00% on Recall@1 after fine-tuning. This exceptional performance demonstrates that EventBind effectively establishes a unified representation space with aligned event, images and text embeddings. To ease future comparison, we report all those numbers in Tab. 5.

References

1. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)

2. Cho, H., Kim, H., Chae, Y., Yoon, K.J.: Label-free event-based object recognition via joint learning with image reconstruction from events. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19866–19877 (2023)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arxiv 2020. arXiv preprint arXiv:2010.11929 (2010)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: 2004 conference on computer vision and pattern recognition workshop. pp. 178–178. IEEE (2004)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
7. Kim, J., Bae, J., Park, G., Zhang, D., Kim, Y.M.: N-imagenet: Towards robust, fine-grained object recognition with event cameras. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2146–2156 (2021)
8. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
9. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
10. Paredes-Vallés, F., De Croon, G.C.: Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3446–3455 (2021)
11. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
12. Posch, C., Matolin, D., Wohlgenannt, R.: A qvga 143 db dynamic range frame-free pwm image sensor with lossless pixel-level video compression and time-domain cds. IEEE Journal of Solid-State Circuits **46**(1), 259–275 (2010)
13. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
14. Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: Events-to-video: Bringing modern computer vision to event cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3857–3866 (2019)
15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
16. Wang, L., Ho, Y.S., Yoon, K.J., et al.: Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10081–10090 (2019)

17. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics **2**, 67–78 (2014)
18. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16816–16825 (2022)