LABEL-FREE EFFICIENT EVENT ENCODER PRETRAINING FROM RGB FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Event cameras offer microsecond latency, high dynamic range, and freedom from motion blur, but labeled event data are scarce and costly to annotate, so event vision has not benefited from foundation-model-scale pretraining. Existing event pretraining methods either learn from events alone at substantial cost (up to hundreds of GPU-hours on large event corpora), or transfer semantics from RGB foundation models by converting events into frames for an image transformer with an appended temporal stage, leaving the recurrent, event-native backbones that are standard in event-based detection without pretraining. We propose a single-stage, label-free cross-modal pretraining method that keeps the event-native recurrent backbone intact. A frozen DINOv3 teacher processes synchronized RGB frames, while an event encoder, a Recurrent Vision Transformer (RVT) followed by DINOv3 transformer blocks, processes the event stream and is trained to match the teacher's global and patch-level features through cosine alignment and an auxiliary patch-level InfoNCE objective. No labels are used, and the RGB branch is discarded after pretraining, so inference is event-only. Pretraining requires 7k steps, 9 A6000 GPU-hours, and 1.2 EFLOPs on DSEC alone, i.e., 2.7× and 26× fewer FLOPs than GEP and TESPEC. On DSEC-Detection, our model reaches 46.3 mAP, improving over RVT trained from scratch (38.4) by +7.9 mAP and over the same architecture trained from scratch (36.6) by +9.7 mAP, with AP_{75} rising from 38.1 to 53.5. Transferring only the pretrained RVT backbone across sensors raises GEN1 from 47.2 to 50.2 mAP, and on DSEC semantic segmentation we obtain 65.68 mIoU, on par with GEP (65.22) and above TESPEC (62.77), both pretrained on substantially larger corpora. These results suggest that dense semantic supervision from an RGB foundation model, applied directly to an event-native backbone, is a compute-efficient alternative to large-scale event pretraining.

1 INTRODUCTION

Event cameras are bio-inspired sensors that report per-pixel brightness changes asynchronously rather than capturing frames (Lichtsteiner et al., 2008; Gallego et al., 2022). This yields microsecond latency, a dynamic range above 120 dB, low power consumption, and no motion blur (Lichtsteiner et al., 2008; Gallego et al., 2022; Gehrig & Scaramuzza, 2024), making them a compelling sensor for fast motion and extreme illumination, precisely where frame-based perception fails.

Exploiting these advantages with modern deep networks, however, is limited by data. Transformers derive much of their strength from large-scale pretraining (Dosovitskiy et al., 2021), and visual foundation models such as CLIP and DINOv3 are trained on hundreds of millions to billions of images (Radford et al., 2021; Siméoni et al., 2025). Event vision has no comparable resource: Perot et al. (2020) list the lack of large-scale datasets among the main reasons for the gap between event- and frame-based systems, and because events are sparse, lack texture, and arrive at very high rates, labeling every timestamp is impractical (Wu et al., 2024). Simulators (Rebecq et al., 2018; Hu et al., 2021) can synthesize events from video, but at the cost of a sim-to-real gap (Stoffregen et al., 2020). Pretraining that requires no event labels is therefore highly desirable.

Two families of label-free event pretraining have emerged, and both are costly. *Event-only self-supervised learning* adapts image pretext tasks to events, from masked modeling (Klenk et al., 2024) and contrastive or dense pretraining (Yang et al., 2023; 2024) to pseudo-intensity reconstruc-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tion (Mohammadi et al., 2025), and requires 76–227 GPU-hours on a single A6000 (Table 4). *Cross-modal methods* import semantics from RGB foundation models but treat events as images: TGVFM reconstructs frames from events for an image model with temporal fusion blocks (Xia et al., 2025), and GEP aligns an image-style ViT to a frozen DINOv2 teacher before a second autoregressive stage on 1.8M event–image pairs (Cao et al., 2026). The recurrent, event-native backbones that are standard in event-based detection, such as RVT (Gehrig & Scaramuzza, 2023), SAST (Peng et al., 2024), and SSM-based models (Zubic et al., 2024), are thus left without semantic pretraining.

We ask whether the semantics of an RGB foundation model can be transferred *directly* into an event-native recurrent backbone, in a single stage and at a fraction of this cost. We propose LEEEP (*Label-free Efficient Event Encoder Pretraining*), which builds on cross-modal distillation (Hinton et al., 2015; Gupta et al., 2016) using synchronized RGB–event pairs (Figure 1): a frozen DINOv3 teacher (Siméoni et al., 2025) encodes the RGB frame, while an RVT trained from scratch, followed by DINOv3 transformer blocks, encodes the event stream and is trained to match the teacher's global and patch-level features through cosine alignment and a patch-level InfoNCE objective. No labels are used, the recurrent backbone is kept intact, and the RGB branch is discarded after pretraining, so inference is event-only.

Pretraining takes 7k steps, 9 GPU-hours, and 1.2 EFLOPs on DSEC alone, 2.7× fewer FLOPs than GEP and 26× fewer than TESPEC. On DSEC-Detection, LEEEP reaches 46.3 mAP, +7.9 over RVT trained from scratch and +9.7 over the same architecture without pretraining (Table 1). Transferring only the pretrained RVT backbone raises GEN1 from 47.2 to 50.2 mAP (Table 2), and on DSEC semantic segmentation it reaches 65.68 mIoU, on par with GEP (65.22) and above TESPEC (62.77) despite their far larger pretraining corpora (Table 3).

Our contributions are:

- LEEEP, a single-stage, label-free cross-modal method that distills a frozen DINOv3 teacher into an event-native recurrent encoder through global and patch-level alignment, with no RGB at inference.
- Evidence that recurrent event backbones can directly absorb foundation-model semantics: +7.9 mAP over RVT on DSEC-Detection and 47.2→50.2 mAP on GEN1 via cross-sensor transfer of the RVT backbone alone.
- A compute analysis showing that 9 GPU-hours of pretraining on DSEC match or exceed methods using 2.7–26× more compute and larger corpora, plus ablations of the encoder, teacher, and each objective (Table 5).

2 Related Work

2.1 Event-Based Visual Backbones

Modern visual recognition has been largely built upon general-purpose image backbones, ranging from convolutional neural networks such as ResNet (He et al., 2016) to transformer-based architectures such as ViT (Dosovitskiy et al., 2021) and Swin Transformer (Liu et al., 2021). These architectures provide strong spatial representation capabilities, but are originally designed for conventional frame-based images and do not explicitly model the sparse, asynchronous, and temporally continuous nature of event streams. This has motivated the development of event-specific architectures that better exploit the characteristics of event data. Early approaches such as AsyNet (Messikommer et al., 2020) leverage event sparsity through asynchronous sparse convolutions, while more recent methods increasingly focus on spatiotemporal representation learning. Recurrent Vision Transformer (RVT) (Gehrig & Scaramuzza, 2023) combines hierarchical transformers with recurrent feature propagation to jointly capture spatial context and temporal dynamics, while SAST (Peng et al., 2024) introduces scene-adaptive sparse attention for efficient event processing, and SSM (Zubic et al., 2024) provides an alternative approach to temporal modeling. Despite these architectural advances, event representations are still predominantly learned through task-specific supervised training, relying on manually annotated event data. As large-scale event annotations remain limited and costly to obtain, recent research has increasingly shifted toward pretraining transferable event representations from unlabeled data.

2.2 Self-Supervised Event Pretraining

Large-scale visual pretraining has demonstrated the effectiveness of learning generalizable representations before downstream adaptation, as exemplified by methods such as MAE (He et al., 2022), CLIP (Radford et al., 2021), and the DINO family of vision foundation models (Caron et al., 2021), particularly the recent DINOv3 (Siméoni et al., 2025). Inspired by the success of representation pretraining in conventional vision, recent studies have extended self-supervised learning to event data, aiming to learn transferable representations without task-specific annotations. MEM (Klenk et al., 2024) first discretizes event representations into visual tokens and then pretrains the event backbone by masking a large portion of the input and predicting the missing discrete tokens, encouraging the encoder to learn contextual and structural information from event streams. ECDDP (Yang et al., 2024) trains the backbone to discover contextual relationships among sparse event patches, encouraging patches with similar contexts to form coherent representations and thereby learning more discriminative event features. TESPEC (Mohammadi et al., 2025) instead reconstructs intensity sequences from masked event inputs, requiring the backbone to preserve both visual appearance and temporal dynamics in its representation. Together, these works demonstrate that self-supervised objectives tailored to event data can guide event backbones toward transferable representations for downstream perception tasks.

2.3 Cross-Modal Event Representation Learning

The methods in 2.2 primarily follow an event-only pretraining paradigm, leaving complementary information from other sensing modalities unexplored. To enrich the learning signal, cross-modal pretraining introduces information from additional modalities. A common choice is paired RGB–event data, which provides naturally synchronized observations and richer appearance and semantic information to guide event representation learning. Early RGB-guided approaches such as ECDP (Yang et al., 2023) transfer knowledge from a pretrained RGB encoder to the event backbone mainly through embedding-level projection and distribution-level alignment. TGVFM (Xia et al., 2025) instead adapts image-based representations to event data through event-to-frame reconstruction and explicit temporal modeling. GEP (Cao et al., 2026) further leverages a frozen vision foundation model in a two-stage framework: it first aligns event representations with image semantics and then performs autoregressive pretraining to capture temporal dependencies in event streams. Our method unifies these objectives within a single pretraining stage. Rather than relying solely on an image-oriented ViT backbone, we retain an event-native RVT front end (Gehrig & Scaramuzza, 2023) to explicitly model event-specific spatiotemporal information. A frozen DINOv3 teacher (Siméoni et al., 2025) further provides semantic guidance through global and patch-level alignment.

3 Method

3.1 Overview

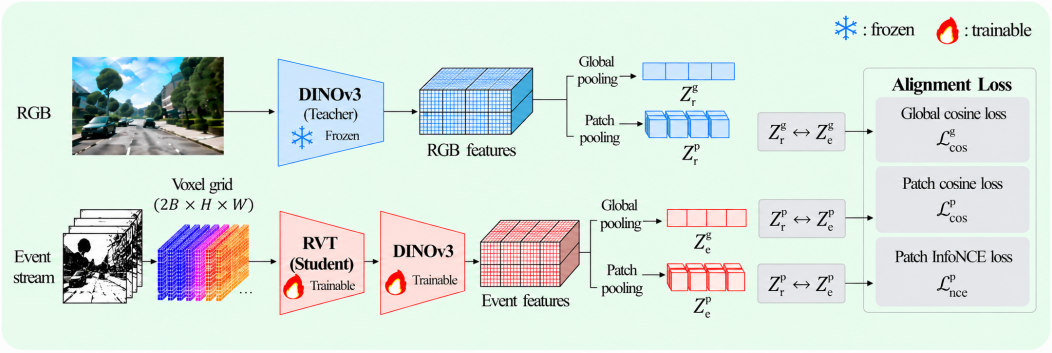

Figure 1: Overview of the proposed RGB-guided event representation pre-training framework. The event model learns from frozen RGB representations through global and patch-level cross-modal alignment.

We propose a single-stage RGB-guided pre-training framework for learning transferable event representations from synchronized RGB–Event observations. Following RVT (Gehrig & Scaramuzza, 2023), the raw event stream is converted into a temporally binned event representation before being processed by the event model. Events within a temporal window Δt are discretized into B temporal bins, with positive and negative polarities accumulated separately, resulting in an event tensor $\mathbf{E}_t \in \mathbb{R}^{2B\times H\times W}$, where H and W denote the spatial resolution.

As illustrated in Figure 1, the framework consists of a frozen RGB teacher and a trainable event model. The RGB teacher employs a pre-trained DINOv3 encoder (Siméoni et al., 2025) to provide high-level semantic supervision from conventional RGB frames. The event model combines RVT with DINOv3 transformer blocks, integrating event-specific spatiotemporal modeling with high-level semantic representation learning.

Given synchronized RGB–Event pairs, we align the event representations with their RGB counterparts at both global and patch levels. Global alignment encourages semantic consistency between the two modalities, while patch-level alignment preserves spatially localized correspondence. An auxiliary patch-level InfoNCE objective further promotes discriminative local representations. During pre-training, the RGB teacher remains frozen and only the event model is optimized. After pre-training, the RGB branch is discarded, and the learned event model is transferred to downstream event perception tasks.

3.2 Cross-Modal Representation Learning

Given a synchronized RGB–Event sample, the RGB branch processes a single RGB frame I_t, while the event branch takes a sequence of T consecutive event representations $\{\mathbf{E}_{t-T+1}, \ldots, \mathbf{E}_t\}$, thereby preserving the temporal information inherent in the event stream.

RGB Teacher. We employ a pretrained DINOv3 ViT (Siméoni et al., 2025) as the RGB teacher and keep its parameters frozen throughout pre-training. Given I_t, the RGB frame is processed through the standard DINOv3 forward path, including patch embedding and the complete Transformer encoder. The resulting patch tokens are reshaped into a spatial feature map $\mathbf{F}_t^r \in \mathbb{R}^{D\times H_p\times W_p}$, where D denotes the DINOv3 embedding dimension and $H_p\times W_p$ its native patch grid. The resulting representation provides a fixed semantic target for learning the event representation.

Event Encoder. The event stream is first processed by an RVT backbone (Gehrig & Scaramuzza, 2023), which combines a multi-stage hierarchical architecture with recurrent LSTM modules. Its multi-stage hierarchical architecture captures spatial features at multiple scales, preserving the rich spatial structure of event data. Meanwhile, the recurrent LSTM modules propagate information across consecutive event representations, effectively modeling the temporal dynamics of the event stream. Overall, effectively leveraging both the spatial structure and temporal dynamics of event data is beneficial for learning informative event representations.

To further enhance the semantic representation of the learned event features, we subsequently employ a DINOv3 ViT after the RVT. During pre-training, the RVT is randomly initialized and trained from scratch, whereas DINOv3 is initialized from pretrained weights and jointly fine-tuned with the RVT. To connect the two architectures, we retain the first three stages of RVT, whose Stage-3 feature map matches the spatial resolution of the native DINOv3 patch-token grid. The RVT features are then mapped to the DINOv3 embedding dimension and processed by the DINOv3 Transformer encoder. This produces the event feature map $\mathbf{F}_t^e \in \mathbb{R}^{D\times H_p\times W_p}$, with the same spatial resolution and feature dimension as $\mathbf{F}_t^r$.

Global and Patch Representations. From the RGB and event feature maps $\mathbf{F}_t^r$ and $\mathbf{F}_t^e$, we derive representations at both global and patch levels. For global representation, global average pooling (GAP) is applied over the spatial dimensions:

$$\mathbf{z}_g^m = \mathrm{GAP}(\mathbf{F}_t^m) = \frac{1}{H_pW_p}\sum_{i=1}^{H_p}\sum_{j=1}^{W_p}\mathbf{F}_t^m(:,i,j), \qquad m\in\{r,e\}, \tag{1}$$

where $m \in \{r, e\}$ denotes the RGB or event modality, respectively, and $\mathbf{z}_g^m \in \mathbb{R}^D$ denotes the corresponding global representation.

For patch-level representation, we directly retain the spatial feature maps without spatial pooling, i.e., $\mathbf{Z}_p^r = \mathbf{F}_t^r$ and $\mathbf{Z}_p^e = \mathbf{F}_t^e$. Since the two feature maps have the same spatial resolution, RGB and event features at the same spatial location form corresponding patch representations for patch-level alignment. The global and patch representations are subsequently used to define the pre-training objectives in 3.3.

3.3 Pre-training Objective

We align the event representations with the frozen RGB teacher at both global and patch levels. The pre-training objective consists of a global cosine loss, a patch-level cosine loss, and a patch-level contrastive loss.

Global Alignment. Given the global representations $\mathbf{z}_g^r$ and $\mathbf{z}_g^e$, we minimize their cosine distance:

$$\mathcal{L}_{\cos}^g = \frac{1}{B}\sum_{b=1}^{B}\left(1 - \frac{(\mathbf{z}_{g,b}^e)^\top \mathbf{z}_{g,b}^r}{\|\mathbf{z}_{g,b}^e\|_2 \|\mathbf{z}_{g,b}^r\|_2}\right), \tag{2}$$

where B denotes the batch size. This objective encourages the event model to learn image-level semantic representations consistent with those of the corresponding RGB frames.

Patch-level Alignment. While global alignment captures image-level semantics, patch-level alignment further provides fine-grained supervision by preserving spatial correspondence between the two modalities. We therefore align RGB and event features at corresponding spatial locations to encourage localized semantic representations. Let $\mathbf{z}_{p,b,i,j}^r$ and $\mathbf{z}_{p,b,i,j}^e$ denote the RGB and event patch-level features at spatial location (i, j) of sample b. The patch-level cosine loss is defined as

$$\mathcal{L}_{\cos}^p = \frac{1}{BH_pW_p}\sum_{b=1}^{B}\sum_{i=1}^{H_p}\sum_{j=1}^{W_p}\left(1 - \frac{(\mathbf{z}_{p,b,i,j}^e)^\top \mathbf{z}_{p,b,i,j}^r}{\|\mathbf{z}_{p,b,i,j}^e\|_2 \|\mathbf{z}_{p,b,i,j}^r\|_2}\right). \tag{3}$$

Patch-level Contrastive Learning. Beyond pairwise alignment, we further introduce contrastive learning at the patch level to improve the discriminability of local representations. Inspired by the in-batch contrastive learning strategy of SimCLR (Chen et al., 2020), we employ a bidirectional InfoNCE objective (van den Oord et al., 2018) to contrast corresponding RGB and event features across samples within each mini-batch. At each spatial location, RGB and event features from the same synchronized sample form a positive pair, while features from other samples within the same mini-batch are treated as negatives. We perform contrastive learning in both directions, where (e, r) denotes event-to-RGB matching and (r, e) denotes RGB-to-event matching:

$$\mathcal{L}_{\text{InfoNCE}}^p = -\frac{1}{2BK}\sum_{(a,b)\in\{(e,r),(r,e)\}}\sum_{n=1}^{B}\sum_{k=1}^{K}\log\frac{\exp\Big(s(\mathbf{z}_{p,n,k}^a, \mathbf{z}_{p,n,k}^b)/\tau\Big)}{\sum_{n'=1}^{B}\exp\Big(s(\mathbf{z}_{p,n,k}^a, \mathbf{z}_{p,n',k}^b)/\tau\Big)}, \tag{4}$$

where $K = H_pW_p$ denotes the total number of spatial locations indexed by k, $s(\cdot,\cdot)$ denotes cosine similarity, and τ is the temperature parameter. We restrict InfoNCE to the patch level because different samples within a mini-batch may share similar image-level semantics and thus become false negatives when contrasting global representations.

Overall Objective. The final pre-training objective combines global semantic alignment, patch-level alignment, and patch-level contrastive learning:

$$\mathcal{L} = \lambda_g\mathcal{L}_{\cos}^g + \lambda_p\mathcal{L}_{\cos}^p + \lambda_c\mathcal{L}_{\text{InfoNCE}}^p, \tag{5}$$

where λ_g, λ_p, and λ_c control the contributions of the three objectives, respectively.

4 Experimental Results

4.1 Experimental Setup

We evaluate the learned event representations on two downstream tasks: object detection on DSEC-Detection and GEN1, with the latter assessing cross-dataset transferability, and semantic segmentation on DSEC. We also evaluate the computational efficiency of our framework.

Datasets. For pre-training, we use 47 sequences from DSEC-Detection (Gehrig & Scaramuzza, 2024; Gehrig et al., 2021), which provides synchronized RGB–event streams, with 38 sequences for training and 9 for validation. For object detection, we evaluate on the remaining 13 DSEC-Detection sequences (Gehrig & Scaramuzza, 2024; Gehrig et al., 2021) for in-domain performance and on GEN1 (de Tournemire et al., 2020) for cross-dataset transfer. For semantic segmentation, we use the 11-class annotations of DSEC (Sun et al., 2022; Gehrig et al., 2021) to evaluate transfer across downstream tasks. Detailed pre-training settings and implementation details are provided in Appendix A.

Baselines. We compare our method against three groups of baselines: (i) our downstream models trained from scratch with the same architecture and training protocol, to assess the benefit of pre-training; (ii) task-specific event-based methods for the corresponding detection or segmentation task; and (iii) existing event representation pre-training methods, including state-of-the-art approaches. For existing methods, we report results from the original papers when evaluated on the corresponding benchmark.

4.2 Object Detection

Table 1: Object detection performance on DSEC-Detection in terms of mAP, AP50, and AP75. Results are reported for RVT (Gehrig & Scaramuzza, 2023), SAST (Peng et al., 2024), SSM (Zubic et al., 2024), LEOD (Wu et al., 2024), TGVFM-S (Xia et al., 2025), and DINOv3 (Siméoni et al., 2025). The best and second-best results are highlighted in bold and underlined, respectively.

Methods	Backbone	Learning Strategy	Pretraining Dataset	mAP↑	AP_{50}↑	AP_{75}↑
Training from scratch						
RVT	MaxViT + RNN	From scratch	–	38.4	58.7	41.3
SAST	Sparse Transformer	From scratch	–	38.1	60.1	40.0
SSM	S5-ViT	From scratch	–	38.0	55.2	40.6
Pretrained Models						
LEOD	RVT	Self-training	–	41.1	65.2	43.6
TGVFM-S	ViT-S/16	Cross-modal KD	DSEC	40.0	**72.7**	38.6
DINOv3	ViT-B/16	Image SSL	LVD-1689M	<u>43.5</u>	65.6	<u>50.8</u>
Ours (scratch)	RVT + DINOv3	From scratch	–	36.6	59.7	38.1
Ours	**RVT + DINOv3**	**Cross-modal SSL**	**DSEC**	**46.3**	<u>70.0</u>	**53.5**

We first assess the effectiveness of our pre-training on in-domain object detection using DSEC-Detection. As shown in Table 1, our method achieves the best overall performance among the directly comparable methods, reaching 46.3 mAP.

Compared with event-based detectors trained from scratch, our method improves upon the strongest baseline, RVT (Gehrig & Scaramuzza, 2023), by 7.9 mAP. It also outperforms methods employing pre-training or enhanced training strategies, exceeding LEOD (Wu et al., 2024) and the label-free TGVFM-S variant (Xia et al., 2025) by 5.2 and 6.3 mAP, respectively. TGVFM-S additionally reports a higher result of 47.7 mAP when label supervision is introduced during pre-training (Xia et al., 2025). Since our work focuses on label-free representation pre-training, we exclude this supervised variant from the direct comparison in Table 1 and report the label-free setting instead.

To isolate the contribution of the proposed cross-modal pre-training, we compare against the same RVT+DINOv3 architecture without cross-modal pre-training. Under the same downstream architecture and training protocol, our pre-training improves mAP from 36.6 to 46.3, corresponding to a 9.7-point gain. This controlled comparison demonstrates that the performance improvement is attributable to the proposed cross-modal pre-training rather than differences in the downstream architecture or training protocol.

Table 2: Object detection performance on GEN1 in terms of mAP. Results are reported for ASTM-Net (Li et al., 2022a), RVT (Gehrig & Scaramuzza, 2023), GET-T (Peng et al., 2023), Swin-T (Liu et al., 2021), ECDDP (Yang et al., 2024), MoBY (Xie et al., 2021), and TESPEC (Mohammadi et al., 2025). The best and second-best results are highlighted in bold and underlined, respectively.

Methods	Backbone	Learning Strategy	Pre-training Dataset	mAP↑
Training from scratch				
ASTMNet	CNN + RNN	From scratch	–	46.7
RVT (baseline)	MaxViT + RNN	From scratch	–	47.2
GET-T	Transformer + RNN	From scratch	–	47.9
Swin-T	Swin-T/7 + RNN	From scratch	–	48.5
Pretrained Models				
ECDDP	Swin-T/7 + RNN	Event SSL	E-TartanAir	49.6
MoBY	Swin-T/7 + RNN	Image SSL	ImageNet-1K	<u>50.2</u>
TESPEC	Swin-T/7 + RNN	Event SSL	1Mpx	**51.6**
Ours	RVT + DINOv3	Cross-modal SSL	DSEC	45.3
Ours	**RVT**	**Cross-modal SSL**	**DSEC**	<u>50.2</u>

We further evaluate the cross-dataset transferability of the learned representations by pre-training on DSEC and fine-tuning for object detection on GEN1. As shown in Table 2, transferring the pre-trained RVT backbone achieves 50.2 mAP, outperforming all models trained from scratch and approaching the best-performing pretrained method, TESPEC (Mohammadi et al., 2025) (51.6 mAP).

Notably, retaining the full RVT+DINOv3 architecture during GEN1 fine-tuning results in 45.3 mAP, whereas transferring only the pretrained RVT backbone improves performance to 50.2 mAP. This suggests that the event-native RVT representations learned through cross-modal pre-training transfer more effectively across datasets than the full multi-modal architecture. A possible explanation is that the image-oriented DINOv3 branch is more sensitive to the domain and spatial-resolution shift between DSEC and GEN1, whereas the RVT backbone retains event-specific spatiotemporal representations that generalize more robustly across datasets.

4.3 Semantic Segmentation

Beyond object detection, we assess the transferability of the learned representations to semantic segmentation on DSEC. As shown in Table 3, our method achieves the best mIoU of 65.68 among the compared methods, together with an mAcc of 74.04.

Compared with models trained from scratch, our method improves over the strongest baseline, ResNet50 (He et al., 2016), by 8.08 mIoU. Among image-based SSL methods, our method exceeds the strongest baseline, DINOv3 (Siméoni et al., 2025), by 4.81 mIoU. Compared with event-specific and cross-modal pre-training methods, our method outperforms TESPEC (Mohammadi et al., 2025) by 2.91 mIoU and achieves comparable performance to GEP (Cao et al., 2026), with a 0.46-point higher mIoU while reaching a slightly lower mAcc (74.04 vs. 74.66). These results demonstrate that the learned event representations transfer effectively to semantic segmentation.

A similar trend is observed in the controlled comparison with the model without cross-modal pre-training. Our pre-training increases mIoU from 46.68 to 65.68 and mAcc from 55.12 to 74.04, corresponding to gains of 19.00 and 18.92 points, respectively. Consistent with the detection results, these improvements further demonstrate the effectiveness of the proposed cross-modal pre-training across downstream tasks.

Table 3: Semantic segmentation performance on DSEC-Seg in terms of mIoU and mAcc. Results are reported for ViT (Dosovitskiy et al., 2021), ResNet (He et al., 2016), MAE (He et al., 2022), BEiT (Bao et al., 2022), DINOv3 (Siméoni et al., 2025), MEM (Klenk et al., 2024), ECDP (Yang et al., 2023), ECDDP (Yang et al., 2024), TESPEC (Mohammadi et al., 2025), and GEP (Cao et al., 2026). The best and second-best results are highlighted in bold and underlined, respectively.

Methods	**Backbone**	**Learning Strategy**	**Pre-training Dataset**	**mIoU↑**	**mAcc↑**
Training from scratch					
ViT	ViT-S/16	From scratch	–	32.66	40.84
ResNet	ResNet50	From scratch	–	57.60	–
Pretrained Models					
MAE	ViT-B/16	Image SSL	ImageNet-1K	51.96	59.84
BEiT	ViT-B/16	Image SSL	ImageNet-1K	51.90	59.66
DINOv3	ViT-B/16	Image SSL	LVD-1689M	60.87	69.92
MEM	ViT-S/16	Event SSL	N-ImageNet	44.62	51.39
ECDP	ViT-S/16	Cross-modal SSL	N-ImageNet	52.52	60.55
ECDDP	ViT-S/16	Event SSL	E-TartanAir	56.38	66.00
TESPEC	Swin-T	Event SSL	1Mpx	62.77	70.61
GEP	ViT-S/14	Cross-modal SSL	Event-1.8M	<u>65.22</u>	**74.66**
Ours (scratch)	RVT + DINOv3	From scratch	–	46.68	55.12
Ours	**RVT + DINOv3**	**Cross-modal SSL**	**DSEC**	**65.68**	<u>74.04</u>

4.4 Computational Efficiency

Our method achieves efficient pre-training with only 9 GPU-hours and 1.2 EFLOPs, as summarized in Table 4. Compared with GEP (Cao et al., 2026), the most efficient competing pre-training method in the table, our method reduces pre-training time from 23 to 9 GPU-hours (60.9%) and computational cost from 3.2 to 1.2 EFLOPs (62.5%). The efficiency advantage remains when considering the complete training pipeline: including DSEC-Seg fine-tuning, our method requires 11.4 GPU-hours in total, compared with 23.3 GPU-hours for GEP (Cao et al., 2026) and 246.4 GPU-hours for TESPEC (Mohammadi et al., 2025), corresponding to reductions of 51.1% and 95.4%, respectively. Within our framework, the pretrained initialization also reduces DSEC-Seg fine-tuning time from 4.8 to 2.4 GPU-hours, corresponding to a 50% reduction. Overall, these results demonstrate the computational efficiency of our method across both pre-training and downstream adaptation.

Table 4: Computational cost comparison for pre-training and DSEC-Seg fine-tuning on a single NVIDIA A6000 GPU. Results are reported for MEM (Klenk et al., 2024), ECDP (Yang et al., 2023), ECDDP (Yang et al., 2024), TESPEC (Mohammadi et al., 2025), and GEP (Cao et al., 2026). Lower values are better, with the best results highlighted in bold.

	Pre-training			**DSEC-Seg. Fine-tuning**		
Method	**Schedule**	**Time (h)↓**	**EFLOPs↓**	**Time (h)↓**	**Params. (M)**	**Total Time (h)↓**
MEM	75 epochs	194	26.5	–	22	–
ECDP	300 epochs	76	10.4	–	21	–
ECDDP	300 epochs	202	27.6	–	21	–
TESPEC	400k steps	227	31.7	19.4	45	246.4
GEP	24 epochs	23	3.2	**0.3**	22	23.3
Ours (scratch)	–	–	–	4.8	89	–
Ours	**7k steps**	**9**	**1.2**	2.4	89	**11.4**

4.5 Ablation Studies

We conduct ablation studies on three key components of our framework: the event encoder architecture, the RGB teacher, and the pre-training objectives. The results are summarized in Table 5.

Table 5: Ablation study of the event encoder, RGB teacher, and pre-training objectives on DSEC object detection. The evaluated architectures include RVT (Gehrig & Scaramuzza, 2023), DINOv3 (Siméoni et al., 2025), DINOv2 (Oquab et al., 2024), and ViTDet (Li et al., 2022b). The best results are highlighted in bold.

Setting No.	RGB Teacher	Event Encoder	Patch	Global	InfoNCE	DSEC-Det mAP↑
Event Encoder						
(a)	DINOv3	**RVT**	✓	✓	✓	38.8
(b)	DINOv3	**DINOv3**	✓	✓	✓	44.2
RGB Teacher						
(c)	**ViTDet**	RVT+DINOv3	✓	✓	✓	40.2
(d)	**DINOv2**	RVT+DINOv3	✓	✓	✓	42.6
Pretraining Objectives						
(e)	DINOv3	RVT+DINOv3	✓	–	–	43.7
(f)	DINOv3	RVT+DINOv3	✓	✓	–	45.4
(g) (Ours)	**DINOv3**	**RVT+DINOv3**	✓	✓	✓	**46.3**

Event Encoder. Under the same cross-modal pre-training scheme, using RVT or DINOv3 alone as the event encoder achieves 38.8 and 44.2 mAP, respectively, while combining RVT with DINOv3 reaches 46.3 mAP, improving over the two variants by 7.5 and 2.1 points. Note that the RVT-only and DINOv3-only variants in this ablation undergo our cross-modal pre-training and therefore differ from the corresponding standalone baselines reported in Table 1, where RVT is trained from scratch and DINOv3 is directly fine-tuned without our cross-modal pre-training. This suggests that the event-native RVT and image-based DINOv3 representations provide complementary benefits for event representation learning.

RGB Teacher. DINOv3 achieves 46.3 mAP, outperforming DINOv2 (42.6) and ViTDet (40.2) by 3.7 and 6.1 points, respectively. These results confirm DINOv3 as the most effective RGB teacher among the evaluated models.

Pre-training Objectives. Patch-level alignment alone achieves 43.7 mAP. Adding global alignment increases performance to 45.4 mAP (+1.7), while incorporating patch-level InfoNCE further improves it to 46.3 mAP (+0.9). The full objective thus achieves the best downstream performance.

5 Conclusion

In this work, we introduced LEEEP, a single-stage RGB-guided pre-training framework for learning transferable event representations. Our event model combines an RVT backbone with DINOv3 transformer blocks and learns from a frozen DINOv3 RGB teacher through global and patch-level alignment and patch-level InfoNCE. The resulting representation achieves 46.3 mAP on DSEC-Detection, transfers across datasets to reach 50.2 mAP on GEN1, and achieves 65.68 mIoU on DSEC semantic segmentation, demonstrating effective transfer across both datasets and downstream tasks. In addition to strong downstream performance, our framework is computationally efficient, requiring only 7k pre-training steps, 9 GPU-hours, and 1.2 EFLOPs on a single NVIDIA A6000 GPU. Overall, these results show that semantic knowledge from RGB foundation models can be efficiently transferred to event-native representations without labels during pre-training.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on experimental design and methodology and to assist with the interpretation and presentation of experimental results. We did not use generative AI tools to generate synthetic datasets, develop theoretical models, formulate mathematical claims, assist with proofs, or implement the proposed method; the remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to assist with literature review and information retrieval, brainstorming, scientific figure editing, and drafting and improving the readability and structure of the manuscript.

All AI-assisted outputs were critically reviewed by the authors. Suggestions related to methodology and experimental design were independently evaluated through experiments, literature-related information was verified against the original sources, and all reported results, interpretations, and conclusions were checked by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed pretraining framework, training objectives, and downstream evaluation protocols in the main paper. The experimental settings, including datasets, data preprocessing, model architectures, optimization configurations, and evaluation metrics, are specified in the corresponding sections and Appendix A. Code and configuration files will be released upon publication to facilitate reproduction of the reported results.

REFERENCES

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022.

Jianwen Cao, Jiaxu Xing, Nico Messikommer, and Davide Scaramuzza. Generative event pretraining with foundation model alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 3189–3199, 2026.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 1597–1607. PMLR, 2020.

Pierre de Tournemire, Davide Nitti, Etienne Perot, Davide Migliore, and Amos Sironi. A large scale event-based detection dataset for automotive. *arXiv preprint arXiv:2001.08499*, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jorg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2022. doi: 10.1109/TPAMI.2020.3008413.

Daniel Gehrig and Davide Scaramuzza. Low-latency automotive vision with event cameras. *Nature*, 629(8014):1034–1040, 2024. doi: 10.1038/s41586-024-07409-w.

Mathias Gehrig and Davide Scaramuzza. Recurrent vision transformers for object detection with event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13884–13893, 2023.

Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021. doi: 10.1109/LRA.2021.3068942.

Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2827–2836, 2016. doi: 10.1109/CVPR.2016.309.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic DVS events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1312–1321, 2021.

Simon Klenk, David Bonello, Lukas Koestler, Nikita Araslanov, and Daniel Cremers. Masked event modeling: Self-supervised pretraining for event cameras. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2378–2388, 2024.

Jianing Li, Siwei Dong, Zhaofei Yu, Yonghong Tian, and Tiejun Huang. Asynchronous spatio-temporal memory network for continuous event-based object detection. *IEEE Transactions on Image Processing*, 31:2975–2987, 2022a.

Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, pp. 280–296, 2022b.

Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120 db 15 μs latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43(2):566–576, 2008. doi: 10.1109/JSSC.2007.914337.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Nico Messikommer, Daniel Gehrig, Antonio Loquercio, and Davide Scaramuzza. Event-based asynchronous sparse convolutional networks. In *European Conference on Computer Vision*, 2020.

Mohammad Mohammadi, Ziyi Wu, and Igor Gilitschenski. TESPEC: Temporally-enhanced self-supervised pretraining for event cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7782–7793, 2025. doi: 10.1109/ICCV51701.2025.00730.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt.

Yansong Peng, Yueyi Zhang, Zhiwei Xiong, Xiaoyan Sun, and Feng Wu. Get: Group event transformer for event-based vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6038–6048, 2023.

Yansong Peng, Hebei Li, Yueyi Zhang, Xiaoyan Sun, and Feng Wu. Scene adaptive sparse transformer for event-based object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16794–16804, 2024. doi: 10.48550/arXiv.2404.01882.

Etienne Perot, Pierre de Tournemire, Davide Nitti, Jonathan Masci, and Amos Sironi. Learning to detect objects with a 1 megapixel event camera. In *Advances in Neural Information Processing Systems*, volume 33, pp. 16639–16652, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. ESIM: an open event camera simulator. In *Proceedings of the 2nd Conference on Robot Learning (CoRL)*, volume 87 of *Proceedings of Machine Learning Research*, pp. 969–982. PMLR, 2018.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Timo Stoffregen, Cedric Scheerlinck, Davide Scaramuzza, Tom Drummond, Nick Barnes, Lindsay Kleeman, and Robert Mahony. Reducing the sim-to-real gap for event cameras. In *European Conference on Computer Vision (ECCV)*, volume 12372 of *Lecture Notes in Computer Science*, pp. 534–549. Springer, 2020. doi: 10.1007/978-3-030-58583-9_32.

Zhaoning Sun, Nico Messikommer, Daniel Gehrig, and Davide Scaramuzza. Ess: Learning event-based semantic segmentation from still images. In *European Conference on Computer Vision (ECCV)*, 2022.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ziyi Wu, Mathias Gehrig, Qing Lyu, Xudong Liu, and Igor Gilitschenski. LEOD: Label-efficient object detection for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16933–16943, 2024.

Ruihao Xia, Junhong Cai, Luziwei Leng, Liuyi Wang, Chengju Liu, Ran Cheng, Yang Tang, and Pan Zhou. Temporal-guided visual foundation models for event-based vision. *arXiv preprint arXiv:2511.06238*, 2025.

Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, and Han Hu. Self-supervised learning with swin transformers. *arXiv preprint arXiv:2105.04553*, 2021.

Yan Yang, Liyuan Pan, and Liu Liu. Event camera data pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10699–10709, 2023.

Yan Yang, Liyuan Pan, and Liu Liu. Event camera data dense pre-training. In *European Conference on Computer Vision (ECCV)*, volume 15101 of *Lecture Notes in Computer Science*, pp. 292–310. Springer, 2024. doi: 10.1007/978-3-031-72775-7_17.

Nikola Zubic, Mathias Gehrig, and Davide Scaramuzza. State space models for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5819–5828, 2024. doi: 10.48550/arXiv.2402.15584.

APPENDIX

A EXPERIMENTAL DETAILS

Pre-training Settings. We construct event representations over a temporal window of 50 ms with 10 bins per polarity, resulting in 20-channel inputs. Each training sample contains a sequence of five event representations. We set the loss weights λ_g, λ_p, and λ_{NCE} to 0.1, 1.0, and 0.1, respectively.

The model is pre-trained for 7k steps with a batch size of 8 using AdamW and a learning rate of 1×10^{-4}, with 1k warmup steps followed by cosine decay. All pre-training experiments are conducted on a single NVIDIA A6000 GPU.

B Attention Visualization

To provide a qualitative analysis of the learned event representations, we visualize the attention maps of the event encoder with and without cross-modal pre-training in Figure 2. For each sample, we show the event input, attention responses without pre-training and with our pre-training, ground-truth annotations, and the corresponding RGB frame. As observed in the figure, our pretrained model produces more object-relevant and less noisy attention responses, with stronger activations around semantically meaningful structures and annotated object regions. In contrast, the model without pre-training exhibits more scattered and background-dominated activations that are less consistently aligned with object regions, particularly for small objects and scenes with challenging illumination. These qualitative results suggest that RGB-guided pre-training helps suppress irrelevant responses and enables the event encoder to capture more discriminative and semantically meaningful representations from event data, consistent with the improvements observed in downstream object detection.

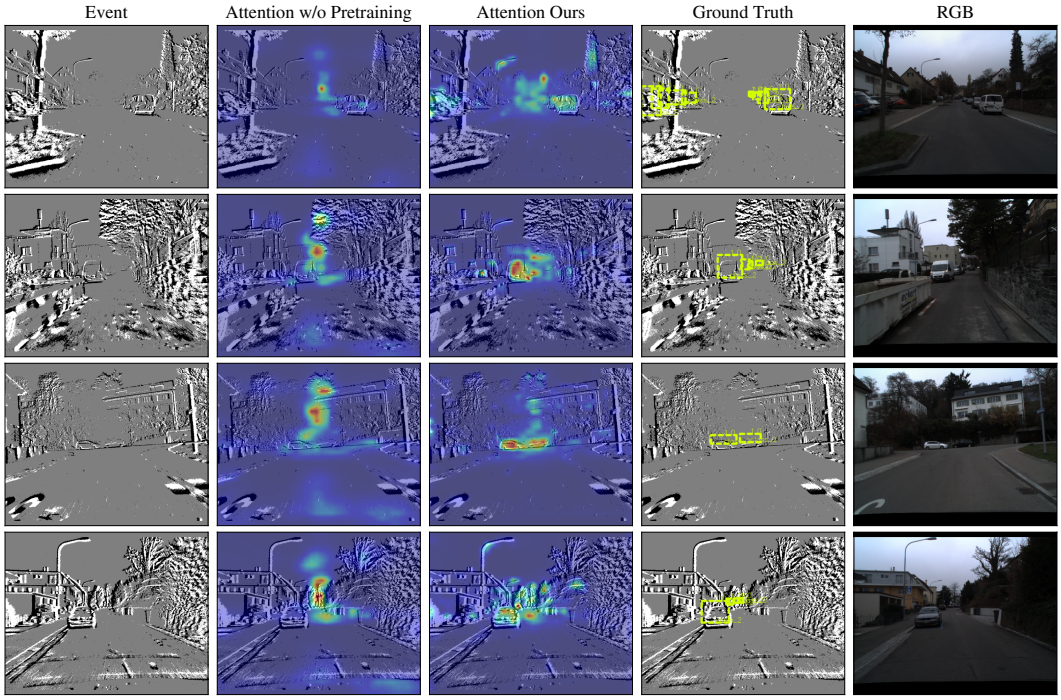

Figure 2: Attention visualization of the event encoder with and without cross-modal pre-training. From left to right: event input, attention without pre-training, attention with our pre-training, ground-truth annotations, and the corresponding RGB frame.

C Source Code

We provide the anonymized source code in a zip file.