Generative Event Pretraining with Foundation Model Alignment

Jianwen Cao Jiaxu Xing Nico Messikommer Davide Scaramuzza

Robotics and Perception Group, University of Zurich

Abstract

Event cameras provide robust visual signals under fast motion and challenging illumination thanks to their microsecond latency and high dynamic range. However, their unique sensing characteristics and limited labeled data make it challenging to train event-based visual foundation models (VFMs), which are crucial for learning visual features transferable across tasks. To tackle this problem, we propose GEP (Generative Event Pretraining), a two-stage framework that transfers semantic knowledge learned from internet-scale image datasets to event data while learning event-specific temporal dynamics. First, an event encoder is aligned to a frozen VFM through a joint regression-contrastive objective, grounding event features in image semantics. Second, a transformer backbone is autoregressively pretrained on mixed event–image sequences to capture the temporal structure unique to events. Our approach outperforms state-of-the-art event pretraining methods on a diverse range of downstream tasks, including object recognition, segmentation, and depth estimation. Together, VFM-guided alignment and generative sequence modeling yield a semantically rich, temporally aware event model that generalizes robustly across domains.

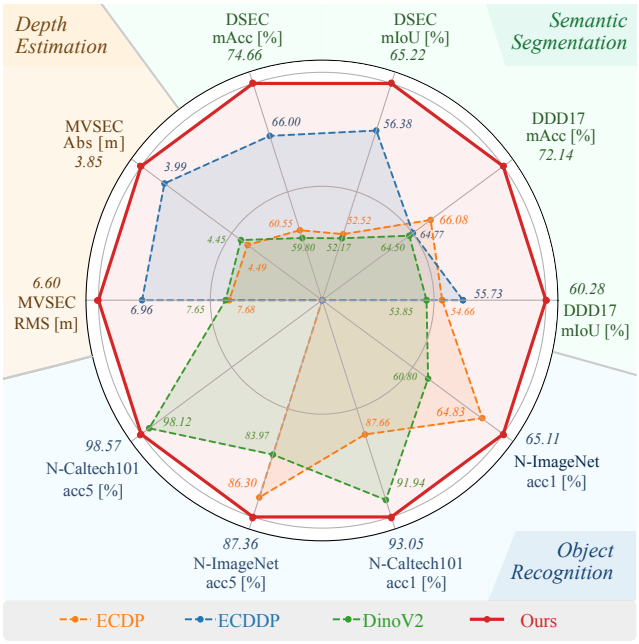


Figure 1. Overall comparison across ten metrics on various datasets. With the same backbone model, our method demonstrates superior and consistent performance across all tasks.

1. Introduction

Event cameras [26] measure per-pixel brightness changes asynchronously with microsecond latency and high dynamic range. Unlike conventional RGB frames, event streams are sparse and provide high temporal resolution, allowing robust perception in challenging lighting and fast-motion scenarios [14]. However, these advantages also introduce challenges: events contain limited texture, differ fundamentally from images, and lack access to large-scale training datasets [14, 37]. These challenges make it difficult to transfer semantic knowledge from images to events, and hinder the robustness of models that rely on large-scale, texture-rich supervision.

Consequently, many existing approaches rely on limited event-only pretraining, small task-specific datasets, or transfer directly from image-pretrained models that are not optimized for the sparse and asynchronous nature of event data [24, 29, 41, 47, 48]. These strategies either miss long-range temporal dynamics or fail to import rich semantic priors from RGB due to suboptimal cross-modality adaptation.

While recent advances in vision foundation models (VFMs) [11, 12, 32, 36] demonstrate that large-scale pretraining on diverse image corpora produces semantically rich and task-transferable representations, these benefits have not yet been fully exploited in the event domain. We address this gap with a unified pretraining paradigm that transfers VFM-level semantics to event data while explicitly modeling temporal dynamics in event streams. By aligning event features with foundation-level image representations, the event encoder inherits broad semantic priors and narrows the discrepancy between event and image modal-

ities. The subsequent generative pretraining stage further equips the model with long-horizon predictive capabilities, enabling scalable, task-agnostic learning on large datasets featuring events and images.

To this end, we introduce *Generative Event Pretraining (GEP)*, a two-stage training framework. In the first stage, an event encoder is aligned to a frozen image encoder (DINOv2 [32]) using a combination of MSE and InfoNCE [31] objectives, transferring semantic structure from the image to the event domain. In the second stage, we perform autoregressive pretraining on aligned event and image sequences using a causal transformer. Unlike masked autoencoding [21], our model predicts future token sequences over randomly masked temporal gaps, encouraging long-horizon temporal reasoning and providing dense generative supervision. We pretrain on the large-scale Event-1.8M corpus with 1.8M event-image pairs, including EventScape [16], N-ImageNet [23], and DSEC [17], covering diverse spatial and temporal patterns.

To evaluate generalization, we conduct experiments on object recognition (N-ImageNet [23], N-Caltech101 [33]), semantic segmentation (DDD17 [3], DSEC [17]), and depth estimation (MVSEC [52]). Among them, N-Caltech101, DDD17, and MVSEC are not included during pretraining and thus serve as out-of-distribution benchmarks.

As summarized in Fig. 1, our method consistently achieves superior performance across all 10 metrics, surpassing both event-specific and image-pretrained baselines while requiring only 24 pretraining epochs. These results demonstrate that aligning event and image features, coupled with autoregressive generative pretraining, enables effective cross-modal transfer and robust task generalization

The main contributions of our work are as follows:

1. *VFM-guided semantic alignment.* We align an event encoder with a pretrained VFM encoder to transfer semantic knowledge learned from large-scale image datasets, effectively grounding event features in rich visual priors.
2. *Autoregressive event pretraining.* We propose an autoregressive pretraining strategy on unlabeled event sequences paired with aligned images, which enables long-horizon temporal reasoning and captures the distinct temporal dynamics of event streams.
3. *Cross-domain generalization.* With about only 15% training epochs compared to previous methods, our model outperforms state-of-the-art methods on recognition (N-ImageNet, N-Caltech101), segmentation (DDD17, DSEC), and depth estimation (MVSEC) benchmarks. Our pretrained backbone demonstrates robust generalization across diverse domains, providing a strong foundation for future event-based vision research.

2. Related Work

2.1. Transfer Learning Across Modalities

In contrast to the conventional pretrain–finetune paradigm in the event domain, transferring knowledge across modalities involves aligning representations between images and events. Prior work in this direction can be broadly divided into two groups: Unsupervised domain adaptation and feature-level distillation.

Unsupervised domain adaptation. UDA bridges the gap between labeled images and unlabeled events by enforcing consistency in features or predictions [29, 41, 46]. ESS [41] aligns embeddings through reconstruction and prediction losses to transfer labels from Cityscapes to unpaired event data. CMESS [46] adds attention-guided soft alignment and joint decoder constraints for dense prediction.

Feature level distillation. Another line distills intermediate features from image models into event encoders using aligned image-event pairs. Hu et al. [22] use grafting by replacing early image layers with an event front end and regressing internal features for label free adaptation. Depth AnyEvent [2] distills multi scale representations into event-based depth estimators and shows that semantic priors help reconstruction and estimation.

Across UDA and distillation, existing methods are designed for a single task and trained on limited data. In contrast, we align events to foundation-level image features in a task-agnostic way and couple the alignment with generative autoregressive pretraining, which captures temporal structure and enables transfer to both recognition and dense prediction tasks.

2.2. Pretraining for Event Vision

Event pretraining targets the learning of general feature representations that are subsequently finetuned for downstream tasks using task labels.

Masked and self reconstruction pretraining. Masked Event Modeling adapts masked autoencoding and frame reconstruction to event streams. MEM [24] follows MAE [21] by masking patches and reconstructing from context. These methods capture local statistics but remain reconstruction-oriented and lack explicit temporal reasoning.

Contrastive and multimodal alignment. Recent methods align events with images or with joint vision–language embedding spaces using contrastive objectives. ECDP [47] aligns event and RGB patch embeddings through momentum-based contrastive learning, leveraging an image encoder pretrained with MoCov3 [7] for supervision but without fully exploiting its rich semantic priors. ECDDP [48] extends this concept to spatial feature maps for pixel level tasks. EventCLIP [44] projects event features into CLIP space for zero-shot and few-shot recognition. EventBind [51] applies multi-stage contrastive finetuning

for stronger event image text alignment. In contrast, our framework adds feature regression toward foundation-level image representations and couples alignment with autoregressive pretraining, which yields semantically grounded and temporally aware event features.

Autoregressive and predictive pretraining. Temporal modeling learns motion dynamics directly from events. Event Transformer [40] introduces a sparse-aware transformer with patch-based representations and latent memory tokens for online recognition. Recent models replace recurrent blocks with structured state space modules S4 and S5 [18, 53], which provide learnable time scales and adapt linear attention language modeling, such as RWKV [34], for asynchronous encoding with multi-step or next representation prediction [20]. These approaches confirm the strong temporal structure of events but are usually single modality, task-specific, and are often trained from scratch on limited data. We pretrain in a task-agnostic way under guidance from vision foundation models to unify semantic alignment with autoregressive temporal reasoning.

2.3. Event VLMs with LLM backbones

A recent popular direction toward generalizable event understanding is to extend large multimodal language models to event data for open-ended reasoning and description. EventGPT [27] aggregates spatiotemporal tokens before feeding a language model to support captioning and question answering. Event-VL [25] builds a generative event-based multimodal model with dynamic semantic alignment. EP VLM [35] uses event prior guided sparsification to prune redundant tokens and can adapt large backbones such as Qwen2-VL [43]. Many event language and video language systems compress visual tokens by about an order of magnitude for efficiency, for example, EventGPT, Event-VL, VideoLLaMA3 [50], and ARVideo [38]. We observe that compression improves reasoning efficiency but harms dense prediction that requires fine spatial detail, see Sec. 4.6. Similar findings have been reported for segmentation and other dense tasks [28, 42]. Large LLM backbones of billions of parameters also hinder real-time inference at high event rates. Our approach avoids heavy language backbones, preserves spatial detail, and transfers to recognition and dense prediction.

3. Method

3.1. Overview

Our method consists of two main stages: *alignment* and *pretraining*. An overview of the entire framework is illustrated in Fig. 2.

Following the event accumulation as detailed in Sec 3.2, each temporal window yields a normalized event frame X_e paired with a synchronized image X_i, as illustrated in Fig. 2 (a). Given a paired event–image sample (X_e, X_i), the event encoder E_e and the image encoder E_i extract modality-specific embeddings (Z_e, Z_i) used for the modality alignment. For the pretraining, we alternately accumulate events and sample images over time, feeding them into the corresponding encoders to obtain a multimodal feature sequence. All encoded embeddings are concatenated into a long sequence S, corresponding to the token sequence fed to the causal transformer in Fig. 2 (b). For each slice $S_{s,s+w}$, a causal transformer T autoregressively predicts the next slice $S_{s+1,s+1+w}$:

$$\hat{S}_{s+1,s+1+w} = T\big(S_{s,s+w}\big), \tag{1}$$

where s is a randomly sampled starting index and w is a fixed window length.

3.2. Event Accumulation

Event streams are composed of asynchronous brightness changes represented as tuples (x, y, t, p), where (x, y) denotes pixel coordinates, t the timestamp, and $p \in \{+1, -1\}$ the polarity indicating whether the brightness increased or decreased. To construct a dense tensor representation suitable for transformer-based modeling, we accumulate events within a fixed temporal window Δt into a pseudo-frame $X_e \in \mathbb{R}^{H\times W\times 3}$.

For each event, we locate its corresponding pixel cell (x, y) and increment the first channel count $M_r(x, y)$ if the polarity is positive ($p = +1$), or the third channel count $M_b(x, y)$ if it is negative ($p = -1$). For the second channel, we assign full intensity whenever any event occurs at (x, y) within the temporal window, serving as an activity mask that marks all active pixels regardless of polarity. After processing all events in the window, the accumulated counts are normalized to suppress local activity spikes and maintain a balanced dynamic range.

Let $M_c(x, y)$ denote the raw accumulated count in channel $c \in \{r, b\}$. We compute a normalization scale α_n as the n-th percentile of all pixel values across both channels (by default $n = 99$). Each channel is then clipped and normalized as

$$X_e(x, y, c) = \frac{\min(M_c(x, y), \alpha_n)}{\alpha_n}, \quad c \in \{r, b\}. \tag{2}$$

This percentile-based clipping and normalization ensure robust scaling and prevent a few high-activity pixels from dominating the distribution, resulting in a stable and balanced input representation for subsequent encoding.

At the end of this stage, each event window produces a normalized event frame X_e that can be paired with a synchronized image frame X_i. These paired samples (X_e, X_i) serve as inputs to the following alignment and pretraining stages.

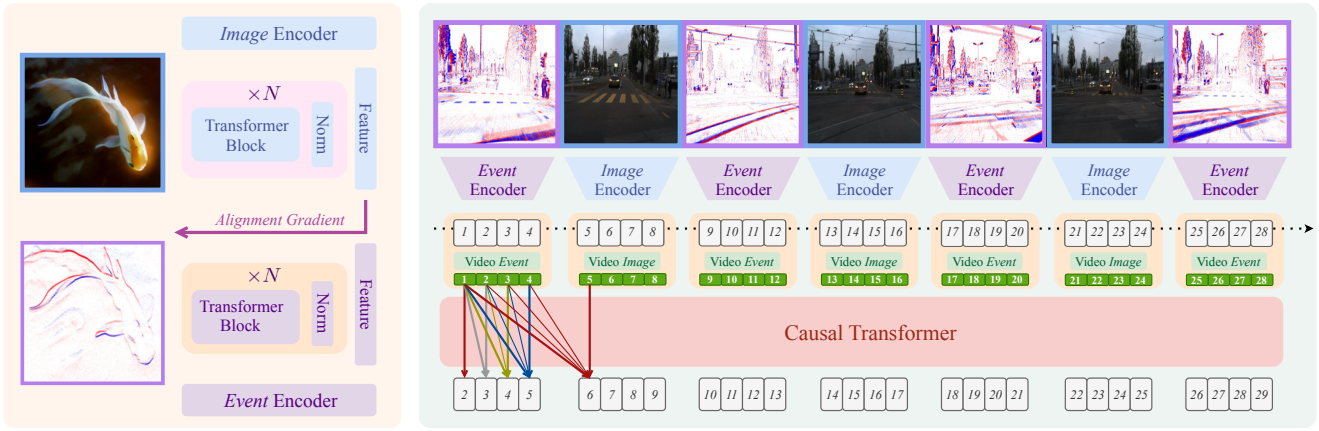


Figure 2. The overall two-stage framework. (a) Alignment stage: Event frames and synchronized images are encoded by an event encoder and a frozen VFM encoder. The event encoder is optimized with regression, contrastive, and preservation terms to match the semantic structure of image features. (b) Autoregressive pretraining stage: Aligned event and image embeddings are interleaved into a single sequence and processed by a causal transformer, which predicts future slices from partial windows, learning long-range temporal structure and cross-modal consistency. Arrows indicate causal dependencies; only a subset is shown for visual clarity.

3.3. Alignment

The weights of both encoders are initialized from DINOv2 [32]. We freeze the image branch E_i (parameters θ_i) and update only the event branch E_e (parameters θ_e). Since vision foundation models already capture strong and well-structured semantics from large-scale RGB corpora, their features provide a reliable reference. Freezing E_i therefore treats the image encoder as a fixed semantic teacher, allowing the event encoder to learn meaningful alignment instead of co-adapting to unstable event features. The alignment objective combines a cosine similarity and an InfoNCE term, while a preservation loss constrains the behavior of E_e on image inputs. Unless otherwise noted, we use DINOv2 [32] as the image foundation model.

Event–image alignment. Given a paired input (X_e, X_i), we extract features

$$Z_e = E_e(X_e), \qquad Z_i = \mathrm{sg}\big(E_i(X_i)\big),$$

where sg stops gradients through the frozen image branch. We use the following alignment loss

$$\mathcal{L}_{\mathrm{a}} = \lambda_{\cos}\left(1 - \cos(Z_e, Z_i)\right) + \lambda_{\mathrm{nce}}\,\mathcal{L}_{\mathrm{nce}}(Z_e, Z_i)\,, \quad (3)$$

where $\cos(\cdot,\cdot)$ denotes cosine similarity. Following standard in-batch contrastive learning, the InfoNCE loss is defined as

$$\mathcal{L}_{\mathrm{nce}} = -\frac{1}{N}\sum_{n=1}^{N}\log\frac{\exp\big(\cos(\tilde{z}_e^{\,n}, \tilde{z}_i^{\,n})/\tau\big)}{\sum_{m=1}^{N}\exp\big(\cos(\tilde{z}_e^{\,n}, \tilde{z}_i^{\,m})/\tau\big)}, \quad (4)$$

where $\tilde{z} = z/\|z\|_2$ denotes ℓ_2-normalized features, τ is the temperature, and N is the batch size. Each positive pair $(\tilde{z}_e^{\,n}, \tilde{z}_i^{\,n})$ is contrasted against all other image embeddings in the batch, which act as negative anchors.

The cosine term enforces directional consistency between event and image embeddings, ensuring that both modalities lie in a shared feature subspace regardless of scale, while the InfoNCE term encourages global separation from unrelated image embeddings, leading to semantically consistent yet discriminative representations. In practice, we apply a lightweight projection head $g(\cdot)$ before (4) following SimCLR [6]. This allows the contrastive branch to adapt its embedding distribution without disturbing the main cosine alignment space, stabilizing optimization and improving cross-modal transfer.

Capability preservation on images. To avoid drift when E_e processes RGB inputs, we pass X_i through E_e and match the frozen image features

$$Z_e^{(I)} = E_e(X_i), \qquad Z_i = \mathrm{sg}\big(E_i(X_i)\big),$$

with the preservation loss

$$\mathcal{L}_{\mathrm{p}} = \mu\left(1 - \cos(Z_e^{(I)}, Z_i)\right). \quad (5)$$

Equation (5) regularizes E_e by preventing its feature space from collapsing or drifting away from the semantic manifold learned by E_i. By enforcing consistent orientation when both branches process image inputs, this term discourages degenerate alignment and maintains semantic grounding.

Total objective. The alignment stage optimizes

$$\mathcal{L} = \mathcal{L}_{\mathrm{a}} + \mathcal{L}_{\mathrm{p}}, \qquad \lambda_{\cos} > 0,\ \lambda_{\mathrm{nce}} > 0,\ \mu > 0. \quad (6)$$

To balance the loss terms, we use separate weights for cosine alignment (λ_{cos}), contrastive separation (λ_{nce}), and preservation (μ). Batch composition and temperature τ follow standard practice for in-batch contrastive learning.

3.4. Pre-training

After alignment, both encoders produce semantically aligned feature embeddings. We concatenate them into a multimodal sequence $S = [S_1, S_2, \ldots, S_K]$ that mixes samples from image datasets and paired event-video datasets, as depicted in Fig. 2 (b). Each token is projected into a shared embedding space and augmented with positional and modality encodings before entering the causal transformer.

Token representation. Given an encoded token S_k, its input representation is

$$X_k = S_k + P_k + M_k, \tag{7}$$

where P_k is the positional embedding and M_k is the modality encoding, indicating whether the token originates from an event frame, image frame, or video sequence. The modality encoding allows the transformer to distinguish heterogeneous sources within the same batch, while positional encoding preserves temporal order.

Dense autoregressive training. We slice a long multimodal sequence into overlapping windows of length w with a stride 1 to form a dense training target. In other words, within one training sample, the model performs

$$S_1 \to S_2, \quad (S_1, S_2) \to S_3, \quad \ldots, \quad S_{1:w-1} \to S_w,$$

yielding a dense sequence of autoregressive targets:

$$\mathcal{L}_{\text{pre}} = \frac{1}{w} \sum_{j=1}^{w} \left\| \hat{S}_{s+j} - S_{s+1+j} \right\|_2^2, \tag{8}$$

encouraging temporally coherent and modality-consistent predictions.

4. Experiments

4.1. Datasets and Baselines

We evaluate on established event-vision benchmarks that cover both recognition, segmentation, and depth estimation tasks. For object recognition, we use N-ImageNet [23] and N-Caltech101 [33], the neuromorphic counterparts of ImageNet [10] and Caltech101 [13]. N-ImageNet contains large-scale event recordings of 1K ImageNet categories captured with a moving DAVIS sensor [4], while N-Caltech101 provides 101 object classes with simpler backgrounds and fewer samples per class. For semantic segmentation, we use DDD17 [3] and DSEC [17, 41], two large-scale driving datasets with synchronized events and grayscale frames. DDD17 offers 12 hours of urban driving sequences with 6 semantic classes derived from intensity frames, while DSEC provides 8,082 higher-resolution stereo event-image-annotation pairs for training. For depth estimation, we use the MVSEC dataset [52], which provides synchronized stereo DAVIS346 event cameras with grayscale frames and accurate depth maps. MVSEC contains indoor and outdoor sequences captured from handheld, hexacopter, car, and motorcycle platforms, enabling depth evaluation under diverse motions and illumination conditions.

For pretraining, we adopt Event-1.8M, a large-scale event corpus that integrates EventScape [16], N-ImageNet, and DSEC. This mixture covers diverse spatial patterns and motion statistics, serving as a task-agnostic dataset for representation learning. DDD17, MVSEC, and N-Caltech101 are excluded from pretraining and used only for out-of-distribution evaluation. More training details are included in the supplementary material.

We compare our model against four groups of baselines to ensure a fair and comprehensive evaluation. The first group trains a ViT [11] from scratch on event data without any pretraining. The second group is specifically designed for each task. The third group includes self-supervised pretraining on images, represented by MAE [21], BeiT [1], and DINOv2 [32]. The last group consists of event-specific pretraining methods, including MEM [24], ECDP [47], ECDDP [48], and EventBind [51].

4.2. Object Recognition

As can be observed by the top-1 and top-5 accuracy reported in Table 1, our model consistently surpasses event-specific pretraining baselines. Compared with ECDP, we improve

Table 1. Object recognition on N-ImageNet and N-Caltech101. We report top-1 (acc1) and top-5 (acc5) accuracies. The two best-performing methods for each evaluation metric are highlighted in green and orange.

Method	Backbone	Dataset	Ep.	N-ImageNet acc1↑	N-ImageNet acc5↑	N-Caltech101 acc1↑	N-Caltech101 acc5↑
Training from scratch							
ViT[11]	ViT-S/16	N-ImageNet	300	46.70	69.89	55.63	–
Specific trained							
EST[15]	–	–	–	48.93	–	–	–
Self-supervised pretraining							
BeiT[1]	ViT-B/16	ImageNet-1K	800	47.15	69.27	53.10	–
MAE[21]	ViT-B/16	ImageNet-1K	800	51.25	72.64	67.68	–
MoCo-v3[7]	ViT-S/16	ImageNet-1K	300	45.77	68.89	76.59	–
DINOv2[32]	ViT-S/16	LVD-142M	–	60.80	83.97	91.94	98.12
Event-specific pretraining							
MEM[24]	ViT-S/16	N-ImageNet	75	57.89	–	–	–
ECDP[47]	ViT-S/16	N-ImageNet	300	64.83	86.30	87.66	–
EventBind[48]	ViT-B/16	N-ImageNet	–	51.40	–	94.08	–
Ours	ViT-S/16	Event-1.8M	24	65.11	87.36	93.05	98.57
Ours	ViT-B/16	Event-1.8M	24	75.20	92.90	96.47	99.56

Table 2. Semantic segmentation on DDD17 and DSEC. We report mean IoU (mIoU) and mean ACC (mAcc). The two best-performing methods for each evaluation metric are highlighted in green and orange.

Method	Backbone	Dataset	Ep.	DDD17		DSEC	
				mIoU↑	mAcc↑	mIoU↑	mAcc↑
Training from scratch							
ViT[11]	ViT-S/16	–	–	36.65	46.21	32.66	40.84
Specific trained							
ESS[41]	–	Cityscape	50	61.37	70.87	53.30	62.94
Self-supervised pretraining							
BeiT[1]	ViT-B/16	ImageNet1K	800	52.39	61.95	51.90	59.66
MAE[21]	ViT-B/16	ImageNet1K	800	53.76	64.78	51.96	59.84
DINOv2[32]	ViT-S/14	LVD-142M	–	53.85	64.50	52.17	59.80
Event-specific pretraining							
MEM[24]	ViT-S/16	N-ImageNet	75	–	–	44.62	51.39
ECDP[47]	ViT-S/16	N-ImageNet	300	54.66	66.08	52.52	60.55
ECDDP[48]	ViT-S/16	E-TartanAir	300	55.73	64.77	56.38	66.00
Ours	ViT-S/14	N-ImageNet	24	58.42	69.59	60.43	68.76
Ours	ViT-S/14	Event-1.8M	24	60.28	72.14	65.22	74.66
Ours	ViT-B/14	Event-1.8M	24	61.90	72.39	67.37	76.27

top-1 from 64.83% to 65.11% on N-ImageNet and from 87.66% to 93.05% on N-Caltech101. Although the event encoder is aligned to DINOv2, our model achieves higher accuracy on event streams than its VFM teacher, trained on 142M RGB images and applied to events. This shows the benefit of the alignment followed by autoregressive modeling to transfer semantics while adapting the features to the sparsity and asynchronicity of events. With only 24 epochs of training (including alignment), our approach outperforms DINOv2 on the tested event benchmarks, demonstrating strong data efficiency and generalization.

4.3. Semantic Segmentation

For DDD17, we use sequence1 for training and the remaining sequences for evaluation, as done in ESS [41]. Following DINOv2, a simple linear segmentation head is attached on top of the encoder. Thus, the performance mainly reflect the quality of the backbone representations. We use cross-entropy and Dice losses [30] on labeled events.

Note that both ECDP and ECDDP adopt a stronger UperNet [45] decoder. ECDP further incorporates a 3D expanded patch embedding [49], while ECDDP additionally applies test-time augmentation with horizontal flipping and multi-scale features. Despite our lighter linear head and the absence of test time augmentation, our method achieves higher segmentation accuracy, indicating that the gains come from the learned representations rather than from a heavy decoder or inference heuristics.

In particular, the variant "Ours, ViT-S/14, N-ImageNet" already surpasses previous baselines on both datasets, with the same pretraining corpus and fewer epochs. When we replace N-ImageNet with the larger Event-1.8M corpus, performance further improves across all metrics, highlighting the importance of large-scale event data.

Table 3. Depth estimation on MVSEC. We report absolute error (Abs↓) and root mean squared error (RMS↓). The two best-performing methods for each evaluation metric are highlighted in green and orange.

Method	Backbone	Dataset	Ep.	Abs↓	RMS↓
Specific trained					
HMNet[19]	–	–	–	4.61	8.60
Self-supervised pretraining					
BeiT[1]	ViT-B/16	ImageNet-1K	800	4.40	7.56
MAE[21]	ViT-B/16	ImageNet-1K	800	4.45	7.60
DINOv2[32]	ViT-S/16	LVD-142M	–	4.45	7.65
Event-specific pretraining					
ECDP[47]	ViT-S/16	N-ImageNet	300	4.49	7.68
ECDDP[48]	ViT-S/16	N-ImageNet	300	3.99	6.96
Ours	ViT-S/16	Event-1.8M	24	3.85	6.60

Table 2 summarizes the semantic segmentation performance reported in mIoU and mAcc. Our model attains the best performance on both datasets, outperforming ECDDP by +4.0 mIoU and +5.0 mAcc on DSEC, indicating that purely static pretraining struggles to capture event dynamics, while our alignment plus autoregressive pretraining bridges this gap. Despite using a fixed DINOv2 as the alignment teacher, the adapted event encoder learns richer spatio-temporal features for dense prediction.

4.4. Depth estimation

We follow the protocol of ECDDP [48] on MVSEC. Training is conducted on the 'outdoor_day2' split, and validation is performed on 'outdoor_day1', 'outdoor_night1', 'outdoor_night2', and 'outdoor_night3'. Following ECDDP, the network is trained to predict normalized log depth. The loss combines a scale-invariant term and a multi-scale gra-

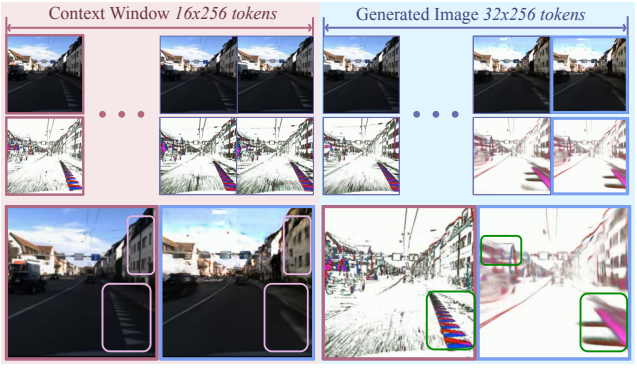


Figure 3. Visualization of a 16-frame context (blue) and a 32-frame autoregressively generated future (red) on validation interleaved event and image streams. Green boxes highlight consistent motion. Insets enlarge the first context and last generated frames for clarity.

(a) Input (b) MEM (c) Ours (d) Label (e) RGB (f) Input (g) ECDDP (h) Ours (i) Label (j) RGB

Figure 4. Qualitative results on the DSEC validation set. The RGB images are shown only for visual reference and are not used by the model.

dient matching term. At test time, predictions are converted back to metric depth, and we report Absolute error and Root Mean Squared error in meters without limiting the maximum depth during evaluation, consistent with prior work [16]. Table 3 summarizes the results. Our method achieves lower Abs and RMS errors than ECDDP and other baselines, demonstrating that the proposed pretraining remains highly effective for depth estimation tasks that require strong 3D spatial understanding.

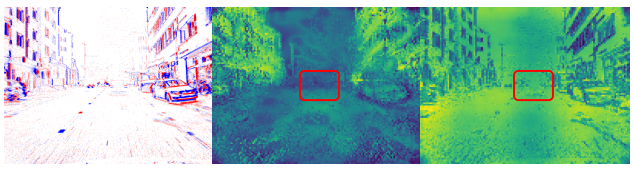
4.5. Qualitative results

Autoregressive long-horizon generation. We visualize sequence rollouts using a frozen, lightweight decoder attached to the encoder, which is used only for qualitative inspection of latent predictions. Each rollout is conditioned on a multimodal context of $256{\times}8{\times}2 = 4096$ tokens and generates $256{\times}16{\times}2 = 8192$ future tokens. Here, 256 denotes spatial tokens per frame, 8 and 16 are the context and predicted steps, and 2 corresponds to interleaved event and image modalities. A sliding causal window is applied when the target horizon exceeds the pretraining window length. Figure 3 indicates that the model extrapolates ego motion, facade parallax, lane-marking drift, and object displacement even when such cues are weak in the last context frame. Despite moderate sensor noise and the minimal decoder, rollouts remain stable and geometrically consistent, which suggests the transformer internalizes long-range temporal structure.

Segmentation results. Figure 4 compares our model with MEM [24] and ECDDP[48]. Our predictions exhibit cleaner boundaries and more coherent region semantics, especially in low-texture or overexposed scenes where event measurements are sparse. In the last row, the roadside pedestrian is barely discernible in the event input due to missing texture, which makes fine structures difficult to detect. With VFM-guided alignment and generative pretraining, the network recovers semantically consistent contours and preserves layout in spite of weak appearance cues.

We refer to the supplementary material for more qualitative task results

Attention responses. Figure 5 analyzes attention on event inputs. The red box marks a distant car that is hard to perceive in the raw events. DINOv2 with registers [8] usually produces meaningful attention on RGB data, yet on events, its response is weak and spatially inconsistent with spurious activations on the road surface. Our aligned and generatively pretrained encoder yields concentrated and structured attention that focuses on the car and other salient elements, indicating effective cross-modal alignment and stronger semantic grounding.



(a) Input image (b) Dinov2 response (c) Our response

Figure 5. Attention map response on event camera data.

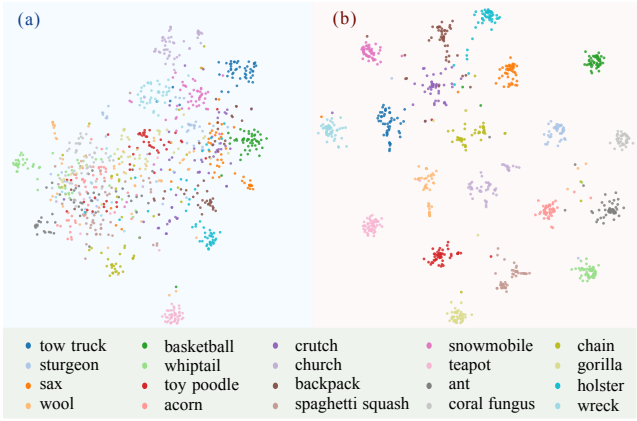


Figure 6. t-SNE visualization of ViT-Base encoder features (20 classes, 50 samples/class from N-ImageNet validation set). Left (a): *Before Align*; right (b): *After Align*. Each color denotes a semantic class.

Table 4. Clustering metrics on encoder features (20 classes, 50 samples/class). Higher is better for Silhouette and Calinski–Harabasz; lower is better for Davies–Bouldin.

	ViT-Small		ViT-Base	
Metric	Dinov2	Aligned	Dinov2	Aligned
Silhouette[39] ↑	-0.03	**0.19**	0.03	**0.27**
Davies–Bouldin[9] ↓	5.24	**2.21**	3.75	**2.04**
Calinski–Harabasz[5] ↑	9.39	**33.21**	14.97	**48.56**

Table 5. Feature alignment results on N-ImageNet using 30,000 training steps. Each cell shows Original→Aligned with relative change (%).

Metric	**Distance Reduction**	**ImageNet Acc. (%)**
MSE	2.22→0.78 (-64.7%)	31.18→61.45
InfoNCE	2.28→0.57 (-75.0%)	31.18→60.99
Cosine	0.70→0.27 (-61.8%)	31.18→61.80
KL Diver.	0.48→0.15 (-67.5%)	31.18→59.20
Attention	38.76→19.42 (-49.9%)	31.18→56.72
Cosine + InfoNCE	1.40→0.36 (-74.3%)	31.18→**62.73**
All	0.96→0.37 (-61.1%)	31.18→61.40

Alignment Clustering. Figure 6 shows a t-SNE plot for event features before and after alignment. Although the alignment pretraining is class-agnostic, it still exhibit pronounced class-wise separation, consistent with near-linear separability in the learned feature space.

Table 4 quantifies the effect, which indicates reduced intra-class variance and stronger inter-class separation, supporting the view that alignment organizes event features into semantically meaningful manifolds that aid downstream tasks.

4.6. Ablations

We report linear probing with a frozen backbone to isolate representation quality in order to ablate alignment, the second-stage autoregressive modeling, masked reconstruction, and token aggregation.

Feature alignment analysis. Table 5 systematically evaluates different alignment objectives on N-ImageNet. Among the individual objectives, *Cosine + InfoNCE* achieves the best trade-off between compactness and transferability, yielding a $+31.6\%$ N-ImageNet accuracy improvement. For the attention-map alignment loss, we match the final-layer attention distributions of the two encoders using an MSE objective. Combining all objectives slightly reduces transfer, suggesting that overly mixed signals may hinder convergence.

Overall, these results confirm that our multi-objective alignment design effectively improves semantic consistency between event and image modalities.

Training paradigm analysis. Removing image-to-event alignment means pretraining on the original DINOv2 feature sequence, which leads to a performance drop on DSEC from 63.27 to 52.38 mIoU, highlighting the importance of alignment. Moreover, the two-stage design outperforms the single-stage variant by +1.30 mIoU, showing that autoregressive pretraining introduces temporal consistency beyond static alignment. In the same setting of our framework, autoregressive (AR) pretraining surpasses masked modeling (MAE) by +0.94 mIoU and +1.48 mAcc, indicating stronger sequence-level reasoning. Even without alignment (*Ours - alignment + finetune*), our autoregressive variant still outperforms state-of-the-art approaches, yielding +5.72 mIoU over ECDP and +1.86 mIoU over ECDDP.

Table 6. Ablation on alignment and training paradigm (frozen backbone).

Variant	Align	Parad.	DSEC (mIoU / mAcc)
Ours (2 stages)	✓	**AR**	**63.27 / 73.13**
Ours - pretrain	✓	None	61.97 / 71.83
Ours - alignment	✗	AR	52.38 / 60.97
Ours - alignment + finetune	✗	AR	58.24 / 66.82
Ours + aggregator [27]	✓	AR	61.24 / 71.53
Masked Modeling + alignment	✓	MAE	62.33 / 71.65

Furthermore, we ablate the downsampling of recent autoregressive video learners and multimodal LLMs, which often cluster spatio-temporal tokens for efficiency [38, 50]. Our controlled pretraining study (*Ours + aggregator*) follows EventGPT [27], which employs a spatio-temporal aggregator that performs independent pooling along the temporal and spatial dimensions and concatenates the results into a compact token sequence. The lower performance achieved by this aggregation demonstrates that aggressive token compression degrades dense prediction, which relies on high spatial fidelity.

5. Conclusion

We presented GEP (Generative Event Pretraining), a unified framework that bridges large-scale VFMs and event-based representation learning. By aligning event features with pretrained VFM representations and performing autoregressive sequence modeling on mixed event–image data, our method effectively transfers semantic knowledge from internet-scale image corpora to the event domain while preserving temporal sensitivity.

Limitations. Our method does not fully account for the asynchronicity of event streams. The autoregressive sequence modeling is implemented with fixed temporal tokenization rather than an arbitrary time-resolution formulation. Future work can build on our framework to develop asynchronous and temporally resolution-agnostic autoregressive models that operate directly on event timestamps.

Acknowledgment

This work was supported by the European Union's Horizon Europe Research and Innovation Programme under grant agreement No. 101120732 (AUTOASSESS) and the European Research Council (ERC) under grant agreement No. 864042 (AGILEFLIGHT).

References

[1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 5, 6

[2] Luca Bartolomei, Enrico Mannocci, Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. Depth anyevent: A cross-modal distillation paradigm for event-based monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19669–19678, 2025. 2

[3] Jonathan Binas, Daniel Neil, Shih-Chii Liu, and Tobi Delbruck. Ddd17: End-to-end davis driving dataset. *arXiv preprint arXiv:1711.01458*, 2017. 2, 5

[4] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014. 5

[5] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974. 8

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020. 4

[7] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021. 2, 5

[8] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023. 7

[9] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, pages 224–227, 2009. 8

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[11] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 5, 6

[12] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19358–19369, 2023. 1

[13] Li Fei-Fei, Rob Fergus, and Pietro Perona. 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*, pages 178–178. IEEE, 2004. 5

[14] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020. 1

[15] Daniel Gehrig, Antonio Loquercio, Konstantinos G Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5633–5643, 2019. 5

[16] Daniel Gehrig, Michelle Rüegg, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction. *IEEE Robotics and Automation Letters*, 6(2):2822–2829, 2021. 2, 5, 7

[17] 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. 2, 5

[18] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021. 3

[19] Ryuhei Hamaguchi, Yasutaka Furukawa, Masaki Onishi, and Ken Sakurada. Hierarchical neural memory network for low latency event processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22867–22876, 2023. 6

[20] Haiqing Hao, Nikola Zubić, Weihua He, Zhipeng Sui, Davide Scaramuzza, and Wenhui Wang. Maximizing asynchronicity in event-based neural networks. *arXiv preprint arXiv:2505.11165*, 2025. 3

[21] 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*, pages 16000–16009, 2022. 2, 5, 6

[22] Yuhuang Hu, Tobi Delbruck, and Shih-Chii Liu. Learning to exploit multiple vision modalities by using grafted networks. In *European Conference on Computer Vision*, pages 85–101. Springer, 2020. 2

[23] Junho Kim, Jaehyeok Bae, Gangin Park, Dongsu Zhang, and Young Min Kim. N-imagenet: Towards robust, fine-grained object recognition with event cameras. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2146–2156, 2021. 2, 5

[24] 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*, pages 2378–2388, 2024. 1, 2, 5, 6, 7

[25] Pengteng Li, Yunfan Lu, Pinghao Song, Wuyang Li, Huizai Yao, and Hui Xiong. Eventvl: Understand event streams

via multimodal large language model. *arXiv preprint arXiv:2501.13707*, 2025. 3

[26] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120 db 15 us latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43 (2):566–576, 2008. 1

[27] Shaoyu Liu, Jianing Li, Guanghui Zhao, Yunjian Zhang, Xin Meng, Fei Richard Yu, Xiangyang Ji, and Ming Li. Eventgpt: Event stream understanding with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29139–29149, 2025. 3, 8

[28] Yifei Liu, Mathias Gehrig, Nico Messikommer, Marco Cannici, and Davide Scaramuzza. Revisiting token pruning for object detection and instance segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2658–2668, 2024. 3

[29] Nico Messikommer, Daniel Gehrig, Mathias Gehrig, and Davide Scaramuzza. Bridging the gap between events and frames through unsupervised domain adaptation. *IEEE Robotics and Automation Letters*, 7(2):3515–3522, 2022. 1, 2

[30] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016. 6

[31] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[32] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 2, 4, 5, 6

[33] Garrick Orchard, Ajinkya Jayawant, Gregory K Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in neuroscience*, 9:437, 2015. 2, 5

[34] Bo Peng, Daniel Goldstein, Quentin Anthony, Alon Albalak, Eric Alcaide, Stella Biderman, Eugene Cheah, Xingjian Du, Teddy Ferdinan, Haowen Hou, et al. Eagle and finch: Rwkv with matrix-valued states and dynamic recurrence. *arXiv preprint arXiv:2404.05892*, 2024. 3

[35] Haotong Qin, Cheng Hu, and Michele Magno. Event-priori-based vision-language model for efficient visual understanding. In *International Joint Conference on Artificial Intelligence*, pages 16–30. Springer, 2025. 3

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1

[37] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):1964–1980, 2019. 1

[38] Sucheng Ren, Hongru Zhu, Chen Wei, Yijiang Li, Alan Yuille, and Cihang Xie. Arvideo: Autoregressive pretraining for self-supervised video representation learning. *arXiv preprint arXiv:2405.15160*, 2024. 3, 8

[39] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987. 8

[40] Alberto Sabater, Luis Montesano, and Ana C Murillo. Event transformer. a sparse-aware solution for efficient event data processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2677–2686, 2022. 3

[41] Zhaoning Sun, Nico Messikommer, Daniel Gehrig, and Davide Scaramuzza. Ess: Learning event-based semantic segmentation from still images. In *European Conference on Computer Vision*, pages 341–357. Springer, 2022. 1, 2, 5, 6

[42] Quan Tang, Bowen Zhang, Jiajun Liu, Fagui Liu, and Yifan Liu. Dynamic token pruning in plain vision transformers for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 777–786, 2023. 3

[43] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3

[44] Ziyi Wu, Xudong Liu, and Igor Gilitschenski. Eventclip: Adapting clip for event-based object recognition. *arXiv preprint arXiv:2306.06354*, 2023. 2

[45] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 6

[46] Chuyun Xie, Wei Gao, and Ren Guo. Cross-modal learning for event-based semantic segmentation via attention soft alignment. *IEEE Robotics and Automation Letters*, 9(3): 2359–2366, 2024. 2

[47] Yan Yang, Liyuan Pan, and Liu Liu. Event camera data pretraining. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2023. 1, 2, 5, 6

[48] Yan Yang, Liyuan Pan, and Liu Liu. Event camera data dense pre-training. In *European Conference on Computer Vision*, pages 292–310. Springer, 2024. 1, 2, 5, 6, 7

[49] Xiaoyu Yue, Shuyang Sun, Zhanghui Kuang, Meng Wei, Philip HS Torr, Wayne Zhang, and Dahua Lin. Vision transformer with progressive sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 387–396, 2021. 6

[50] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 3, 8

[51] Jiazhou Zhou, Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. Eventbind: Learning a unified representation to bind them all for event-based open-world understanding. In *European*

Conference on Computer Vision, pages 477–494. Springer, 2024. 2, 5

[52] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3 (3):2032–2039, 2018. 2, 5

[53] 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*, pages 5819–5828, 2024. 3