LLM-REGPROTO: LLM-ENABLED CROSS-MODAL PROTOTYPE REGULARIZER FOR WEAKLY SUPERVISED EVENT SEMANTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Event cameras offer microsecond temporal resolution and high dynamic range, making them attractive for autonomous perception. However, the sparsity and texture-free nature of event data render pixel-wise annotation prohibitively expensive, bottlenecking fully supervised semantic segmentation. We present LLM-RegProto, a weakly supervised framework requiring only extremely sparse point annotations (as few as 1 class 1 point, or a standard 1 class 10 points) to achieve competitive segmentation. The core innovation repurposes a frozen, pretrained text-only LLM layer as a cross-modal prototype regularizer: visual tokens from the event-stream encoder are projected into a high-dimensional linguistic manifold, where the frozen transformer layer induces semantically structured class prototypes via its pretrained attention mechanism. Unlike multimodal LLM approaches requiring paired vision-language training or text prompts, our framework feeds visual features directly into a pure-text LLM layer, with class prototypes formed autonomously through learnable class query tokens without textual supervision. To bridge the visual–linguistic modality gap, we design (i) a cross-modal adapter with 2D spatial positional encoding and bidirectional attention, enabling the autoregressive LLM backbone to process non-sequential spatial tokens; and (ii) a dual-stream prototype alignment network with background-noise suppression, jointly training two encoders on temporally complementary event windows and filtering ego-motion-induced clutter via contrastive prototype alignment. On DDD17 and DSEC benchmarks under the 1C10C budget, LLM-RegProto achieves mIoU gains of +2.16% and +2.04% over prior state-of-the-art weakly supervised methods, demonstrating that LLM pretraining weights encode universal structured priors transferable across modalities without any fine-tuning.

1 INTRODUCTION

Event cameras respond to per-pixel brightness changes asynchronously, generating streams $\mathcal{E} = \{(x_i, y_i, t_i, p_i)\}$ with microsecond temporal resolution and dynamic range exceeding 140 dB, making them well suited for high-speed and low-light perception tasks Muhammad et al. (2022); Zhao et al. (2023); Gallego et al. (2022). Semantic segmentation, which assigns a semantic label to each pixel, is a fundamental capability for autonomous driving and robotics Cheng et al. (2022); Xie et al. (2021). However, segmenting event data remains severely under-explored, primarily due to a costly annotation bottleneck. Unlike RGB images, event representations are sparse, polarity-coded, and lack color or texture cues Zhu et al. (2019). Dense pixel annotation in this domain is cognitively demanding, time-consuming, and error-prone, severely limiting the availability of labeled training data.

To reduce annotation costs, weakly supervised learning has emerged as a compelling direction Ahn & Kwak (2018); Wang et al. (2020); Xu et al. (2023b). In the RGB domain, point-supervised segmentation has advanced through class activation maps, propagation networks, and prototype matching Kweon & Yoon (2024). Directly transferring these techniques to the event domain, however, faces three fundamental challenges. First, there is no pretrained visual backbone for event data comparable to ImageNet-pretrained models, leaving prototype clustering without a reliable feature foundation. Second, the structural mismatch between event voxel grids and natural image statistics

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.

undermines techniques that rely on spatial cue memorization. Third, ego-motion generates abundant background events that interfere with semantic prototype formation, causing clustering collapse or contamination.

Existing solutions address these challenges only partially. Supervised event segmentation methods require expensive dense annotations Alonso & Murillo (2019); Wang et al. (2021); Sun et al. (2022); Jia et al. (2023); Biswas et al. (2024); Long et al. (2025). Domain adaptation approaches transfer knowledge from RGB sources but suffer from significant domain shift Rebecq et al. (2021); Gehrig et al. (2020); Sun et al. (2022); Jing et al. (2024). Prototype-based weakly supervised methods in the RGB domain Zhou et al. (2022); Xu et al. (2022) typically rely on hand-crafted class-name text prompts or visual descriptors to initialize class centroids, which do not transfer to the event domain where no textual class associations exist. Meanwhile, vision-language models like CLIP Radford et al. (2021) and LLaVA Liu et al. (2023) achieve impressive zero-shot recognition by leveraging paired image-text data, but this requires multimodal pretraining and textual class names—neither of which are available in our setting.

These observations motivate our core idea. The pretrained attention mechanism of a language model Vaswani et al. (2017) encodes a universal semantic clustering geometry that operates on text tokens. We hypothesize that this geometry could also cluster visual tokens projected into the same hidden space, even when the model has never seen any images Pang et al. (2024). In other words, a frozen, text-only LLM could function as a zero-shot semantic regularizer for a completely different modality. We project event-derived visual tokens into the LLM hidden space via a lightweight cross-modal adapter and process them through a single frozen transformer layer. No LLM weights are updated, and no textual class descriptions are provided. Language thus acts as a geometric prior, not a semantic vocabulary.

Building on this insight, we introduce LLM-RegProto, a weakly supervised semantic segmentation framework for event cameras. The main contributions of this paper are as follows:

- We repurpose a frozen text-only LLM as a cross-modal prototype regularizer. A single frozen LLM transformer layer, pretrained on text only with no multimodal alignment, functions as a zero-shot semantic clustering operator without any weight updates. Learnable class query tokens, carrying no pre-assigned class semantics and requiring no textual descriptions or class-name prompts, are processed through the frozen attention mechanism to produce structured per-class prototypes. These prototypes are driven entirely by visual loss signals and evolve through training purely from visual supervision.
- We design a dual-stream prototype alignment network with background-noise suppression. Two parallel event encoders process complementary temporal windows with different ego-motion profiles. A mask-driven exclusion mechanism operates at both the prototype alignment level and the pseudo-label level, preventing ego-motion noise from corrupting the dual-supervision signal and the prototype geometry on the linguistic manifold. This explicit background suppression enables stable prototype learning under challenging real-world conditions.
- In the 1C1C setting, we outperform the existing best weakly supervised method by 1.78%; in the 1C10C setting, we outperform the existing best weakly supervised method by 2.16%, achieving new state-of-the-art results and surpassing several fully supervised baselines.

2 Related Works

2.1 Event Camera Segmentation and Weak Supervision

Given a raw event stream $\mathcal{E} = \{(x_i, y_i, t_i, p_i)\}$, existing methods convert it into fixed-size tensors—event frames, voxel grids, TORE volumes, or time surfaces for downstream processing Gallego et al. (2022); Zhu et al. (2019). The voxel grid $\mathbf{V} \in \mathbb{R}^{T\times H\times W}$ has become the de facto standard for dense prediction tasks due to its compatibility with 3D CNNs.

In the RGB image domain, supervision signals span image-level class tags Ahn & Kwak (2018); Wang et al. (2020), bounding boxes, scribbles, and point clicks. Among these, point supervision is the paradigm most relevant to our work: each annotated point provides a precise spatial location with its class label, making it more informative per annotation than image-level tags but far cheaper than

dense masks Xu et al. (2023b). Critically, existing point-supervised methods rely on color and texture cues for pseudo-label expansion—cues absent in event data. EV-WSSS Cho et al. (2024) pioneered weakly supervised event segmentation, proposing a framework for bidirectional branching to process event data, as well as a task setting and dataset for weakly supervised point annotation. However, it did not address the problem of insufficient information in the event data itself. Our work focuses on point-supervised event segmentation and addresses this gap by replacing texture-based region growing with LLM-regularized prototype alignment and dual-stream temporal consistency.

2.2 Prototype Learning and Contrastive Alignment

Per-class prototypes originate in metric and few-shot learning and have been widely adopted in semi-supervised segmentation Zhou et al. (2022); Xu et al. (2022). ProtoSeg and memory-bank methods aggregate class-representative feature vectors to enable non-parametric nearest-neighbor classification, but they require a substantial fraction of dense labels and become unstable at extremely low label rates. Cross-domain prototype alignment Jiang et al. (2022); Das et al. (2023) aligns source and target prototype distributions for domain adaptive segmentation. We extend this idea to the temporal domain, aligning prototypes across complementary event windows rather than across domains. Foundation model-derived prototypes—via DINOv2 Oquab et al. (2024) or SAM Kirillov et al. (2023) offer strong semantic coherence but rely on visual pretraining. Our approach derives structure from a pretrained LLM's attention geometry, providing a complementary and more annotation-efficient source of semantic priors.

2.3 Language Models as Visual Priors

The convergence of language and vision has driven a new generation of cross-modal representation learners. CLIP Radford et al. (2021) and ALIGN Jia et al. (2021) train dual encoders via contrastive losses for zero-shot transfer, while multimodal models like Flamingo Alayrac et al. (2022) and LLaVA Liu et al. (2023) treat frozen language models as sequence reasoners over visual tokens. For dense prediction, ODISE Xu et al. (2023a) and LISA Lai et al. (2024) extend these representations to pixel-level tasks. Closer to our design, recent methods Tang et al. (2025); Pang et al. (2024); Zhu et al. (2024) employ frozen LLM layers as structural feature transformers, demonstrating that pretrained attention weights generalize across modalities without fine-tuning. Building on this, our approach utilizes a single frozen layer from a strictly text-only LLM as a semantic clustering function. However, unlike CLIP-style encoders or visual instruction models that inherently rely on text-image alignment data, our prototype formation is driven entirely by learnable visual query tokens, requiring no textual prompts, class names, or multimodal pretraining.

3 Methodology

3.1 Framework Overview

An event camera fires an event at pixel (x, y) when the log-intensity change exceeds a threshold θ:

$$\log I(x, y, t) - \log I(x, y, t_{\text{last}}) = p \cdot \theta, \quad p \in \{+1, -1\} \tag{1}$$

where $I(\cdot)$ denotes the pixel intensity and t_{last} is the timestamp of the preceding event at the same pixel. The raw event stream is converted into a voxel grid representation $\mathbf{V} \in \mathbb{R}^{B\times T\times H\times W}$ by temporally binning events:

$$V_k(x, y) = \sum_{i:t_i\in[t_k,t_{k+1}]} p_i \cdot \delta(x - x_i, y - y_i) \tag{2}$$

where B is the batch size, T is the number of temporal bins (typically $T = 5$), and $H \times W$ is the spatial resolution.

Given an event voxel sequence $\mathbf{V}^{(s)} \in \mathbb{R}^{T\times H\times W}$ and a sparse point annotation set $\mathcal{P} = \{(x_j^c, y_j^c, c)\}_{j=1}^{M}$, where $c \in \{1, \ldots, K\}$ is the class label and $M \ll H \times W$ (as few as $M = K \cdot 10$),

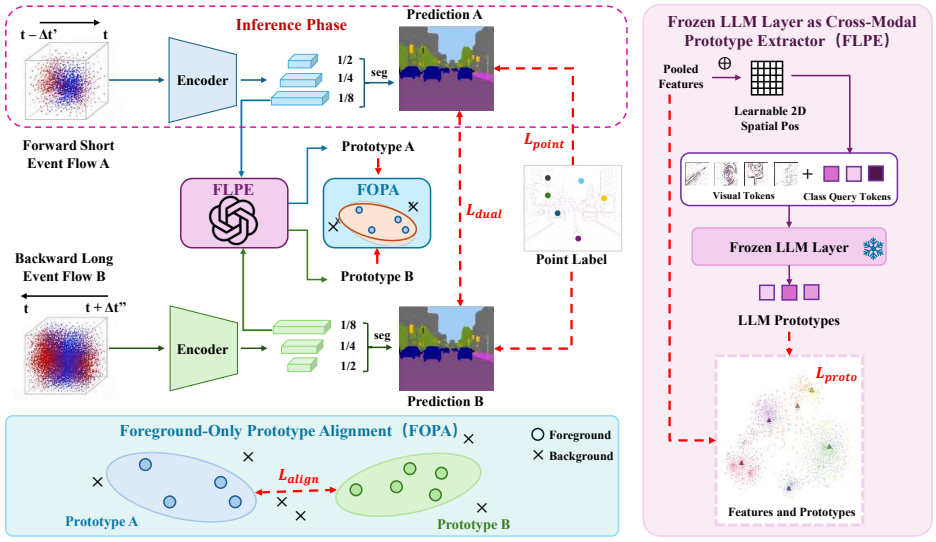


Figure 1: Our model constructs a mutual supervision mechanism through event flows in front and behind time windows to achieve prediction consistency under point-level labeling constraints. Simultaneously, it introduces a frozen large model layer (FLPE) to extract semantic prototypes and combines it with foreground-only prototype alignment (FOPA) to enhance feature discrimination ability.

the goal is to predict a dense segmentation map $\hat{\mathbf{Y}} \in \{1, \ldots, K\}^{H \times W}$. We define the sparse label tensor:

$$\mathbf{L}(x, y) = \begin{cases} c, & \text{if } (x, y, c) \in \mathcal{P} \\ 255, & \text{otherwise (ignore)} \end{cases} \tag{3}$$

where 255 is the standard ignore index.

To address this, LLM-RegProto employs a dual-stream architecture equipped with a frozen-LLM prototyper. This dual-stream design exploits temporal complementarity Cho et al. (2024): Stream A accumulates events forward ($t - \Delta t' \to t$), while Stream B accumulates backward ($t + \Delta t'' \to t$). Although both observe the same target time, their differences in temporal direction, accumulation length ($\Delta t' < \Delta t''$), and ego-motion noise patterns dictate that their encoder weights remain independent. This isolation provides the crucial complementary signals upon which FOPA depends.

Within the architecture, the encoder (e.g., $\mathcal{E}_A$) is a ConvLSTM-UNet that recurrently processes temporal sub-windows of the event voxel grid, producing a hidden state $\mathbf{s}^{(i)}$ and a multi-scale feature pyramid $\mathbf{f}^{(i)} = \{\mathbf{f}_r^{(i)}\}_{r \in \{2,4,8\}}$. The corresponding decoder $\mathcal{D}_A$ then reconstructs dense logits via transposed convolutions and skip connections.

During training, both streams and the frozen-LLM prototyper provide joint supervision. In inference, only Stream A and the prototyper are utilized. The final full-resolution prediction combines the decoder logits $\mathbf{Z}_A$ with the bilinearly upsampled prototype similarity maps $\mathbf{S}^{\uparrow} \in \mathbb{R}^{K \times H \times W}$:

$$\hat{\mathbf{Y}} = \arg\max_k \left(\mathbf{Z}_A^{(k)} + \mathbf{S}^{\uparrow(k)} \right) \tag{4}$$

where the lower-resolution similarity map $\mathbf{S} \in \mathbb{R}^{K \times H_8 \times W_8}$ is scaled by τ_{seg} during training.

3.2 Frozen LLM Layer as Cross-Modal Prototype Extractor

Modern autoregressive large language models (LLMs), pretrained on vast text corpora, inherently develop a highly structured semantic latent space. Formally, let $\mathcal{T} \subset \mathbb{R}^{d_h}$ denote the token embedding

space of a pretrained LLM. As well-documented in probing studies, transformer processing projects semantically related tokens into close geometric proximity within this space. We conceptualize the linguistic manifold, $\mathcal{M}_{\text{lang}}$, as the low-dimensional submanifold of $\mathbb{R}^{d_h}$ spanned by the embeddings of semantically coherent token sequences.

We hypothesize that the transformer attention mechanism of a frozen LLM functions as a universal semantic clustering operator, $\mathcal{F} : \mathbb{R}^{N\times d_h} \to \mathbb{R}^{N\times d_h}$, satisfying the following distance constraints:

$$\|\mathcal{F}(\mathbf{T})_i - \mathcal{F}(\mathbf{T})_j\|_2 < \epsilon \quad \text{if sem}(\mathbf{T}_i) = \text{sem}(\mathbf{T}_j) \tag{5}$$

$$\|\mathcal{F}(\mathbf{T})_i - \mathcal{F}(\mathbf{T})_j\|_2 > \delta \quad \text{if sem}(\mathbf{T}_i) \neq \text{sem}(\mathbf{T}_j) \tag{6}$$

where $\text{sem}(\cdot)$ denotes semantic category membership, and $\epsilon \ll \delta$. Crucially, $\mathcal{F}$ generalizes to arbitrary input tokens—regardless of their native modality—provided they are mapped into a compatible subspace of $\mathbb{R}^{d_h}$. To achieve this, a learned adapter $\phi : \mathbb{R}^D \to \mathbb{R}^{d_h}$ serves as a cross-modal bridge. By projecting event-derived visual features into this space, we align them with the pretrained semantic topology of $\mathcal{F}$. This enables the frozen attention mechanism to naturally induce class-discriminative prototypes without any architectural or parameter updates to the LLM.

This cross-modal generalization is rooted in the transformer's weighted token aggregation. During pretraining, attention weights internalize deep relational priors governing token interactions. Because these priors encode general contextual co-occurrence rather than strict syntax, they are fundamentally semantic. Once visual features are projected into the shared latent space, they readily adopt this contextual structure, explaining why frozen LLM attention exhibits such robust generalization across entirely different modalities.

Cross-Modal Adapter with 2D Spatial Encoding Let $\mathbf{F} \in \mathbb{R}^{B\times D\times H_8\times W_8}$ denote the encoder output at $1/8$ spatial resolution with channel dimensionality $D = 256$. The prototyper module Φ processes this dense feature map in three main steps. First, adaptive average pooling spatially compresses the features to a fixed resolution of $(P_h, P_w) = (16, 16)$, yielding a set of discrete spatial tokens. A learnable 2D positional embedding $\mathbf{P}_{\text{2D}}$ is then added to these tokens to preserve their spatial adjacency. Next, the tokens are linearly projected to match the LLM's hidden dimensionality ($d_h = 2048$). They are concatenated with learnable class query tokens and passed through a single pretrained LLM transformer layer.

To facilitate global context aggregation, we remove the standard causal mask, enabling unconstrained bidirectional attention. The frozen attention mechanism operates over the combined token sequence $\mathbf{X}$ as follows:

$$\text{Attn}(\mathbf{X}) = \text{softmax}\left(\frac{\mathbf{X}\mathbf{W}_Q^{\text{LLM}}(\mathbf{X}\mathbf{W}_K^{\text{LLM}})^T}{\sqrt{d_k}}\right)\mathbf{X}\mathbf{W}_V^{\text{LLM}} \tag{7}$$

The injected $\mathbf{P}_{\text{2D}}$ empowers the LLM to effectively exploit positional proximity during this semantic clustering process. Crucially, to maintain the integrity of the pretrained representational space, all LLM parameters remain strictly frozen throughout training ($\nabla_{\theta_{\text{LLM}}} = \mathbf{0}$).

Prototype Extraction and Supervision After the frozen LLM layer, the output tokens at the class query positions are extracted as the per-class prototypes $\{\mathbf{p}_k\}_{k=1}^K$, each of dimension d_h. Before similarity computation, each prototype vector is L2-normalized to the unit hypersphere. Similarity is computed between each normalized prototype and each LLM-processed spatial token (also L2-normalized), both residing on the linguistic manifold after the frozen LLM layer. The similarity maps supervise annotated points via a cross-entropy loss with temperature scaling:

$$\mathcal{L}_{\text{proto}} = \text{CE}\big(\mathbf{S}^{\uparrow}/\tau_{\text{seg}},\ \mathbf{L}\big) \tag{8}$$

where $\mathbf{L}$ is the sparse point label tensor (ignore index 255) and $\tau_{\text{seg}} = 0.1$ sharpens the similarity distributions. This loss drives the frozen-LLM-induced prototypes to align spatially with annotated class locations.

3.3 FOREGROUND-ONLY PROTOTYPE ALIGNMENT

Background Class Exclusion The background class in event cameras is driven primarily by ego-motion artifacts rather than meaningful semantics. Treating all classes equally allows this noise to distort the prototype geometry. We address this via a dual-level background exclusion mechanism applied during both prototype alignment and pseudo-labeling.

Ego-motion causes background prototypes to diverge across temporal windows for purely non-semantic reasons. Using them as negative samples introduces spurious repulsion in the contrastive space. To prevent this geometric distortion during prototype alignment, we strictly restrict the contrastive loss $\mathcal{L}_{\text{align}}$ to the foreground classes $k \in \mathcal{F} = \{0, \dots, K-1\}$, explicitly masking the background ($k = K$). Concurrently, excluding background pixels when generating cross-stream pseudo-labels ensures this noise does not corrupt the exchanged supervisory signals.

Consequently, the background-excluded contrastive loss over the foreground prototypes $\{\mathbf{p}_{A,i}\}_{i\in\mathcal{F}}$ and $\{\mathbf{p}_{B,i}\}_{i\in\mathcal{F}}$ is formulated as a symmetric cross-entropy objective:

$$\mathcal{L}_{\text{align}} = \frac{1}{2}\Big[\text{CE}(\mathbf{\Sigma}, \mathbf{y}) + \text{CE}(\mathbf{\Sigma}^T, \mathbf{y})\Big] \tag{9}$$

where $\mathbf{\Sigma}_{ij} = \mathbf{p}_{A,i} \cdot \mathbf{p}_{B,j}/\tau_{\text{proto}}$ represents the cross-stream cosine similarity matrix exclusively for the foreground ($i, j \in \mathcal{F}$), and $y_i = i$ constitutes the diagonal ground-truth targets. This symmetric formulation is mathematically equivalent to the bidirectional InfoNCE objective:

$$\mathcal{L}_{\text{align}} = -\frac{1}{2|\mathcal{F}|}\sum_{i\in\mathcal{F}}\left[\log\frac{\exp(\mathbf{p}_{A,i}\cdot\mathbf{p}_{B,i}/\tau_{\text{proto}})}{\sum_{j\in\mathcal{F}}\exp(\mathbf{p}_{A,i}\cdot\mathbf{p}_{B,j}/\tau_{\text{proto}})} + \log\frac{\exp(\mathbf{p}_{A,i}\cdot\mathbf{p}_{B,i}/\tau_{\text{proto}})}{\sum_{j\in\mathcal{F}}\exp(\mathbf{p}_{A,j}\cdot\mathbf{p}_{B,i}/\tau_{\text{proto}})}\right] \tag{10}$$

By explicitly capturing bidirectional prototype matching, this objective maximizes a lower bound on the mutual information $\mathcal{I}(\mathbf{p}_A; \mathbf{p}_B)$ between shared foreground semantic concepts across both streams.

Dual Pseudo-Label Consistency To enforce cross-stream consistency, we derive confidence-thresholded hard pseudo-labels from the softmax predictions of each stream using a confidence threshold of $\tau = 0.5$. Critically, these targets exclude the background class as well as any classes absent from the point annotation set. Letting $\tilde{\mathbf{Y}}_A$ and $\tilde{\mathbf{Y}}_B$ denote the refined pseudo-labels derived from Stream A and Stream B, respectively, the streams bidirectionally supervise one another via the objective:

$$\mathcal{L}_{\text{dual}} = \text{CE}(\mathbf{Z}_A,\ \tilde{\mathbf{Y}}_B) + \text{CE}(\mathbf{Z}_B,\ \tilde{\mathbf{Y}}_A) \tag{11}$$

where CE is the standard pixel-wise cross-entropy loss employing an ignore index of 255. Enforcing background exclusion at this pseudo-label level guarantees that spurious, noise-driven background predictions cannot propagate through the cross-supervisory channels. Finally, to ensure numerical stability during optimization, degenerate pseudo-labels that map entirely to the ignore index are systematically discarded.

Training Objective The overall training objective is formulated as a weighted superposition of the individual loss components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{point}} + \lambda_1\mathcal{L}_{\text{dual}} + \lambda_2\mathcal{L}_{\text{proto}} + \lambda_3\mathcal{L}_{\text{align}} \tag{12}$$

where the hyperparameters scaling the auxiliary objectives are empirically set to $\lambda_1 = 0.5$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.01$.

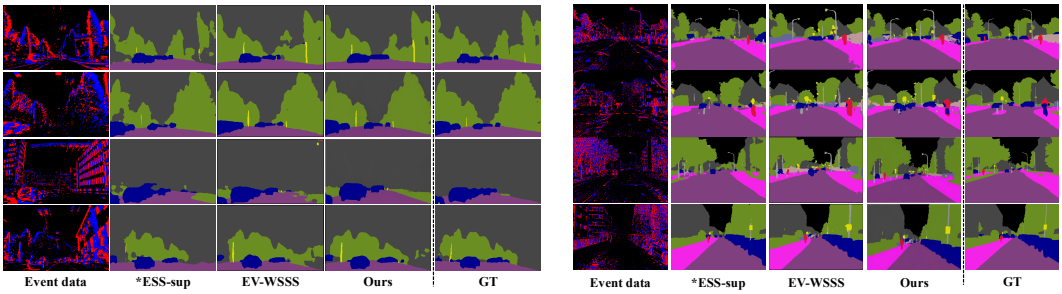
As part of the delayed activation strategy, the model is optimized exclusively with the primary loss for the first two epochs; the auxiliary constraints ($\mathcal{L}_{\text{dual}}$, $\mathcal{L}_{\text{proto}}$, and $\mathcal{L}_{\text{align}}$) are suspended to achieve early structural stability. Crucially, the core LLM parameters are shielded from these updates, remaining entirely frozen ($\nabla_{\theta_{\text{LLM}}} = \mathbf{0}$) throughout the entirety of the optimization process.

Table 1: Comparison with event-based semantic segmentation methods on DDD17 dataset.

Method	Pub.	mIoU
Fully Supervised		
Ev-SegNetAlonso & Murillo (2019)	CVPRW'19	54.81
DTLWang et al. (2021)	ICCV'21	58.80
ESS-supSun et al. (2022)	ECCV'22	61.37
EvSegformerJia et al. (2023)	TIP'23	54.41
HALSIEBiswas et al. (2024)	WACV'24	60.66
Spike-BRGNetLong et al. (2025)	TCSVT'24	54.72
Unsupervised DA		
E2VIDRebecq et al. (2021)	TPAMI'19	48.74
Vid2EGehrig et al. (2020)	CVPR'20	45.48
EV-TransferMessikommer et al. (2022)	RA-L'22	15.52
ESSSun et al. (2022)	ECCV'22	53.09
Weakly Supervised		
EV-WSSS (1C1C)Cho et al. (2024)	ECCV'24	56.88
EV-WSSS (1C10C)Cho et al. (2024)	ECCV'24	60.39
LLM-RegProto (1C1C)	Ours	$\mathbf{58.36 \pm 0.25}$
LLM-RegProto (1C10C)	Ours	$\mathbf{62.55 \pm 0.28}$

Table 2: Comparison with event-based semantic segmentation methods on DSEC dataset.

Method	Pub.	mIoU
Fully Supervised		
Ev-SegNetAlonso & Murillo (2019)	CVPRW'19	51.76
ESS-supSun et al. (2022)	ECCV'22	53.29
HMNet-BHamaguchi et al. (2023)	CVPR'23	51.20
HALSIEBiswas et al. (2024)	WACV'24	52.43
Spike-BRGNetLong et al. (2025)	TCSVT'24	54.95
Unsupervised DA		
E2VIDRebecq et al. (2021)	TPAMI'19	44.08
EV-TransferMessikommer et al. (2022)	RA-L'22	24.37
ESSSun et al. (2022)	ECCV'22	45.38
Weakly Supervised		
EV-WSSS (1C1C)Cho et al. (2024)	ECCV'24	45.55
EV-WSSS (1C10C)Cho et al. (2024)	ECCV'24	51.19
LLM-RegProto (1C1C)	Ours	$\mathbf{47.33 \pm 0.28}$
LLM-RegProto (1C10C)	Ours	$\mathbf{53.23 \pm 0.11}$



(a) DDD17 **(b) DSEC**

Figure 2: Qualitative results on DDD17 and DSEC. * denotes full supervision.

4 Experiments

4.1 Datasets

DDD17 is a large-scale event-based driving dataset recorded with a DAVIS camera (346×200), providing synchronized event streams and grayscale frames with semantic labels for 6 classes: flat, background, object, vegetation, human, vehicle. Following, we use the official train/val/test splits, with 15,950 training frames and 3,890 test frames.

DSEC-Semantic is a high-resolution (640×440) stereo event camera dataset collected in urban driving scenarios, with 11 semantic classes. We use the official semantic segmentation split with 8,082 training and 2,809 test frames.

4.2 Implementation Details

For evaluation, the study employed two sparse point annotation settings: 1 class 10 points (denoted as 1C10C) and 1 class 1 point (denoted as 1C1C) per frame. Regarding model architecture and training details, the model uses a UNet-ConvLSTM with 3 encoder blocks and skip connections as the encoder, paired with a lightweight convolutional decoder, and employs a single-layer frozen Llama-3.2-1B and DeepSeek-R1-Distill-Qwen-1.5B as the LLM backbone. The model uses the RAdam optimizer with a learning rate of 2×10^{-4} and a batch size of 4. The training process was configured with specific loss weights and temperature parameters and was completed entirely on a single NVIDIA 4090 (24GB) GPU.

4.3 Comparison with State-of-the-Art

Existing event segmentation methods follow three main paradigms: fully supervised learning Alonso & Murillo (2019); Wang et al. (2021); Sun et al. (2022); Jia et al. (2023); Biswas et al. (2024); Long et al. (2025), domain adaptation Rebecq et al. (2021); Gehrig et al. (2020); Sun et al. (2022), and weak

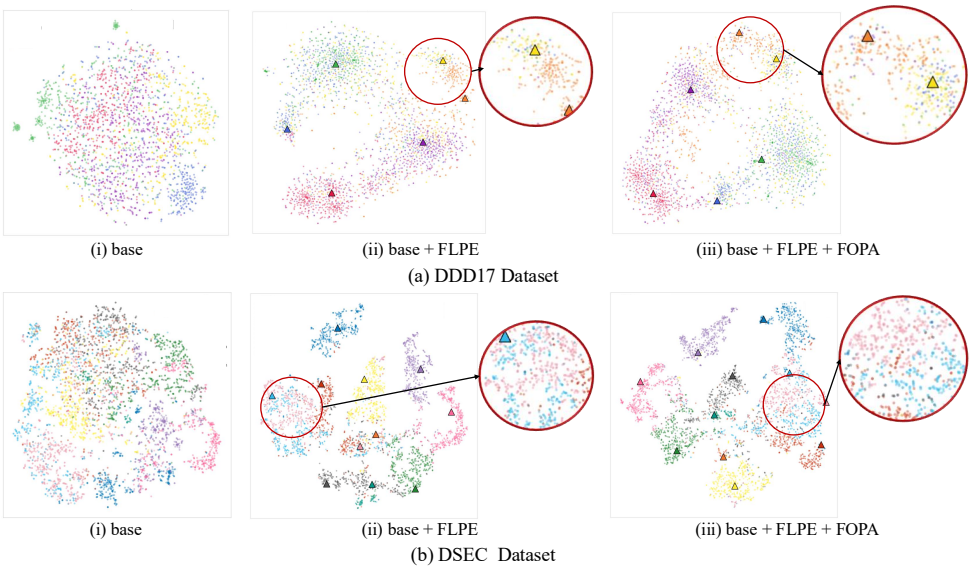


Figure 3: t-SNE visualization of per-class prototype distributions.

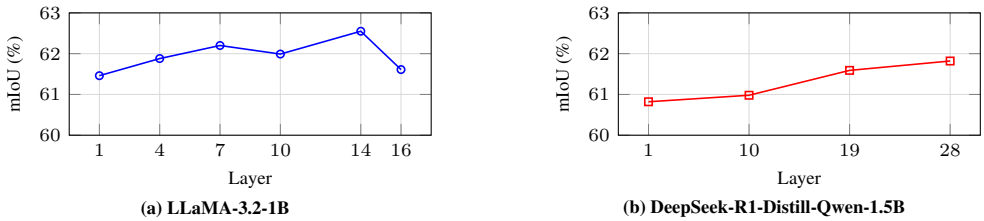


Figure 4: Effect of frozen LLM layer depth on DDD17 under the 1C10C setting.

supervision Cho et al. (2024). Fully supervised methods require expensive pixel-wise annotations, while domain adaptation reduces target-domain labeling but remains affected by RGB-to-event domain shift. Point-supervised methods are more annotation-efficient, yet must propagate extremely sparse labels without color or texture cues. LLM-RegProto addresses this challenge using frozen LLM representations, without dense labels or paired source data.

As detailed in Figure 2 and Tables 1 and 2, LLM-RegProto establishes a new state of the art across both benchmarks and annotation budgets. Under 1C10C, our method obtains 62.55 ± 0.28 mIoU on DDD17 and 53.23 ± 0.11 on DSEC, outperforming EV-WSSS by 2.16 and 2.04 points, respectively. Under the extreme 1C1C constraint, it obtains 58.36 ± 0.25 and 47.33 ± 0.28, corresponding to gains of 1.48 and 1.78 points. The improvements exceed the observed training-seed variation.

4.4 Prototype Geometry Analysis

Figure 3 visualizes the learned prototype distributions via t-SNE on both DDD17 Binas et al. (2017) and DSEC Gehrig et al. (2021). On DDD17, the dominant effect of FLPE is a substantial enlargement of inter-class distances: the Base model produces overlapping clusters, while the full model yields maximally separated, tightly bounded class regions. On DSEC, the improvement manifests primarily as sharper decision boundaries between adjacent clusters, with transition zones narrowing and intra-class density increasing—particularly for visually similar categories. Together, the two benchmarks confirm that FLPE organizes visual prototypes onto the linguistic manifold in a geometrically consistent manner across domains.

4.5 Ablation Study

LLM layer selection Figure 4 ablates the choice of LLM backbone and the frozen layer used for prototype regularization. For LLaMA-3.2-1B (a 16-layer model), performance follows a non-

monotonic trajectory: it improves from the 1st layer to a peak at the 14th (62.55), before declining at the final 16th layer. This aligns with the well-known inverted U-shaped performance curve of transformers in transfer tasks: early layers capture low-level features, middle-to-late layers develop transferable semantic structures, and final layers often collapse toward task-specific representations.

Situated near the network's end, the 14th layer of LLaMA-3.2-1B retains the richest semantic clustering geometry while remaining generic enough for cross-modal transfer, explaining its superiority. Conversely, DeepSeek-R1-Distill-Qwen-1.5B (a 28-layer reasoning-distilled model) shows monotonic improvement with depth, peaking at the final layer (61.82) but still falling short of LLaMA-3.2-1B's maximum. This suggests that its distillation process shifted internal representations toward specialized reasoning, rendering them less effective as generic semantic clusterers for visual tokens. Consequently, we adopt LLaMA-3.2-1B as the default backbone.

Controlled validation of LLM pretraining To separate the effect of pretraining from Transformer capacity, we replace the frozen Llama-3.2-1B layer with architecture-matched random and weight-shuffled controls while leaving the remaining framework, including FOPA, unchanged. The random controls match the pretrained layer in architecture and parameter count, while weight shuffling destroys learned organization but preserves marginal values.

Table 3: Architecture-matched controls on DDD17 (1C10C). "Random init." denotes an architecture-matched randomly initialized Transformer; all non-Base variants retain FOPA.

Model variant	mIoU (%)
Base (w/o FLPE/FOPA)	58.29
Random init. (frozen)	60.41
Shuffled pretrained weights (frozen)	60.53
Random init. (trainable)	60.97
Pretrained weights (frozen)	**62.55**

As shown in Table 3, all controls trail the pretrained frozen layer by 1.58 to 2.14 points. Since FOPA is retained, their gains over the Base cannot be attributed solely to the inserted Transformer: the FOPA-only variant already reaches 60.34 mIoU in Table 8. Relative to this result, the random frozen block adds only 0.07 points, whereas pretrained weights raise performance from 60.41 to 62.55. Thus, the improvement primarily reflects pretrained interaction structure rather than attention capacity alone.

Component design ablation Table 4 details each component's contribution under the 1C10C annotation budget. The dual-stream ConvLSTM-UNet Base achieves 58.29 mIoU on DDD17 and 51.03 on DSEC. Integrating the Frozen LLM Prototype Extractor (FLPE) yields substantial gains (+3.46 and +1.37 mIoU, respectively). This demonstrates that a frozen LLM's pretrained attention geometry can induce semantically structured prototypes directly from visual tokens, requiring neither weight updates nor textual supervision.

Adding Foreground-Only Prototype Alignment (FOPA) further raises performance to 62.55 and 53.23 mIoU, gains of 0.80 and 0.83 points. By aligning foreground prototypes across streams while masking the background, FOPA prevents ego-motion noise from distorting the FLPE-induced geometry. The total gains over the Base (4.26 and 2.20 points) confirm their complementary roles: FLPE supplies semantic structure, while FOPA stabilizes it across temporal views.

Table 4: Ablation of key components (1C10C).

Base	FLPE	FOPA	mIoU (%) DDD17	mIoU (%) DSEC
✓			58.29	51.03
✓	✓		61.75	52.40
✓	✓	✓	**62.55**	**53.23**

5 Conclusion

We presented LLM-RegProto, a weakly supervised event camera segmentation framework that reuses a single frozen plain text LLM layer as a zero-shot cross-modal prototyping regularizer. Using one frozen text-only LLM layer to regularize class prototypes for point-supervised event segmentation. Its 2D adapter and bidirectional attention transfer pretrained token interactions without text, multimodal pretraining, or LLM updates, while dual-stream foreground alignment suppresses ego-motion noise. New state-of-the-art results on DDD17 and DSEC under both 1C1C and 1C10C demonstrate the value of frozen language-model structure for event representations.

6 LLM USAGE STATEMENT

Large Language Models (LLMs) were used to aid in the writing and polishing of the manuscript. Specifically, we used an LLM to assist in refining the language, improving readability, and ensuring clarity in various sections of the paper and assist in locating relevant research. The model helped with tasks such as sentence rephrasing, grammar checking, and enhancing the overall flow of the text.

It is important to note that the LLM was not involved in the ideation, research methodology, or experimental design. All research concepts, ideas, and analyses were developed and conducted by the authors. The contributions of the LLM were solely focused on improving the linguistic quality of the paper, with no involvement in the scientific content or data analysis.

The authors take full responsibility for the content of the manuscript, including any text generated or polished by the LLM. We have ensured that the LLM-generated text adheres to ethical guidelines and does not contribute to plagiarism or scientific misconduct.

REFERENCES

Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022.

Iñigo Alonso and Ana C. Murillo. Ev-segnet: Semantic segmentation for event-based cameras. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1624–1633, 2019. doi: 10.1109/CVPRW.2019.00205.

Jonathan Binas, Daniel Neil, Shih-Chii Liu, and Tobi Delbruck. DDD17: end-to-end DAVIS driving dataset. *CoRR*, abs/1711.01458, 2017. URL `http://arxiv.org/abs/1711.01458`.

Shristi Das Biswas, Adarsh Kosta, Chamika Liyanagedera, Marco Apolinario, and Kaushik Roy. Halsie: Hybrid approach to learning segmentation by simultaneously exploiting image and event modalities. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5952–5962, 2024. doi: 10.1109/WACV57701.2024.00586.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Garg. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1290–1299, 2022.

Hoonhee Cho, Sung-Hoon Yoon, Hyeokjun Kweon, and Kuk-Jin Yoon. Finding meaning in points: Weakly supervised semantic segmentation for event cameras. In *European Conference on Computer Vision*, pp. 266–286. Springer, 2024.

Anurag Das, Yongqin Xian, Dengxin Dai, and Bernt Schiele. Weakly-supervised domain adaptive semantic segmentation with prototypical contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15434–15443, June 2023.

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, 2022.

Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3583–3592, 2020. doi: 10.1109/CVPR42600.2020.00364.

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.

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*, pp. 22867–22876, June 2023.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.

Zexi Jia, Kaichao You, Weihua He, Yang Tian, Yongxiang Feng, Yaoyuan Wang, Xu Jia, Yihang Lou, Jingyi Zhang, Guoqi Li, and Ziyang Zhang. Event-based semantic segmentation with posterior attention. *IEEE Transactions on Image Processing*, 32:1829–1842, 2023. doi: 10.1109/TIP.2023.3249579.

Zhengkai Jiang, Yuxi Li, Ceyuan Yang, Peng Gao, Yabiao Wang, Ying Tai, and Chengjie Wang. Prototypical contrast adaptation for domain adaptive semantic segmentation. In *Computer Vision - ECCV 2022*, volume 13694 of *Lecture Notes in Computer Science*, pp. 36–54. Springer, 2022. doi: 10.1007/978-3-031-19830-4_3.

Linglin Jing, Yiming Ding, Yunpeng Gao, Zhigang Wang, Xu Yan, Dong Wang, Gerald Schaefer, Hui Fang, Bin Zhao, and Xuelong Li. Hpl-ess: Hybrid pseudo-labeling for unsupervised event-based semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23128–23137, June 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Hyeokjun Kweon and Kuk-Jin Yoon. From SAM to CAMs: Exploring segment anything model for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3599–3608, 2024.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9579–9589, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Xianlei Long, Xiaxin Zhu, Fangming Guo, Chao Chen, Xiangwei Zhu, Fuqiang Gu, Songyu Yuan, and Chunlong Zhang. Spike-brgnet: Efficient and accurate event-based semantic segmentation with boundary region-guided spiking neural networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(3), 2025.

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. doi: 10.1109/LRA.2022.3145053.

Khan Muhammad, Tanveer Hussain, Hayat Ullah, Javier Del Ser, Mahdi Rezaei, Neeraj Kumar, Mohammad Hijji, Paolo Bellavista, and Victor Hugo C de Albuquerque. Vision-based semantic segmentation in scene understanding for autonomous driving: Recent achievements, challenges, and outlooks. *IEEE Transactions on Intelligent Transportation Systems*, 23(12):22694–22715, 2022. doi: 10.1109/TITS.2022.3207665.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaa El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Ziqi Pang, Ziyang Xie, Yunze Man, and Yu-Xiong Wang. Frozen transformers in language models are effective visual encoder layers. In *International Conference on Learning Representations*, 2024.

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*, pp. 8748–8763. PMLR, 2021.

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, 2021. doi: 10.1109/TPAMI.2019.2963386.

Zhaoning Sun, Nico Messikommer, Daniel Gehrig, and Davide Scaramuzza. Ess: Learning event-based semantic segmentation from still images. In *European Conference on Computer Vision*, pp. 341–357. Springer, 2022.

Fenghe Tang, Wenxin Ma, Zhiyang He, Xiaodong Tao, Zihang Jiang, and S. Kevin Zhou. Pre-Trained LLM is a Semantic-Aware and Generalizable Segmentation Booster . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15969. Springer Nature Switzerland, September 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Lin Wang, Yujeong Chae, and Kuk-Jin Yoon. Dual transfer learning for event-based end-task prediction via pluggable event to image translation. In *2021 ICCV*, pp. 2115–2125, 2021. doi: 10.1109/ICCV48922.2021.00214.

Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12275–12284, 2020.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12077–12090, 2021.

Haiming Xu, Lingqiao Liu, Qiuchen Bian, and Zhen Yang. Semi-supervised semantic segmentation with prototype-based consistency regularization. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26007–26020. Curran Associates, Inc., 2022. doi: 10.52202/068431-1886.

Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2955–2966, 2023a.

Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Learning multi-modal class-specific tokens for weakly supervised dense object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19596–19605, 2023b.

Rui Zhao, Ruiqin Xiong, Jian Zhang, Zhaofei Yu, Shuyuan Zhu, Lei Ma, and Tiejun Huang. Spike camera image reconstruction using deep spiking neural networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(6):5207–5212, 2023. doi: 10.1109/TCSVT.2023.3326294.

Tianfei Zhou, Wenguan Wang, Ender Konukoglu, and Luc Van Gool. Rethinking semantic segmentation: A prototype view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2582–2593, 2022.

Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 989–997, 2019.

Lei Zhu, Fangyun Wei, and Yanye Lu. Beyond text: Frozen large language models in visual signal comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27047–27057, June 2024. doi: 10.1109/CVPR52733.2024.02554.

Table 5: Sensitivity of prototype contrastive temperature τ on DDD17.

τ	0.1	0.3	0.5	0.7	1
1C1C	56.33	57.45	57.87	58.36	57.13
1C10C	62.55	62.35	61.72	62.22	61.70

Table 6: Ablation of FLPE components on DDD17 (1C10C).

Base	**LLM proto**	**2D pos**	**Suppress causal**	**mIoU (%)**
✓				58.29
✓	✓			60.53
✓	✓	✓		61.37
✓	✓	✓	✓	**61.75**

A Sensitivity to contrastive temperature

Table 5 reveals that the optimal contrastive temperature differs substantially between the two annotation budgets, exposing an important interplay between supervision density and prototype sharpness. Under the 1C10C setting, the optimal temperature is $\tau = 0.1$, which produces a sharp similarity distribution that forces prototypes to be well-separated. With ample annotation points, the class query tokens receive sufficient gradient signals to converge to distinct cluster centers, and a low temperature maximizes inter-class margin without causing overfitting.

Under the 1C1C setting, the optimal temperature shifts to $\tau = 0.7$. At this extremely low annotation budget, each class is anchored by only a single point per image, making the prototypes highly sensitive to annotation noise and stochasticity across images. A higher temperature softens the similarity distribution, allowing prototypes to overlap slightly and preventing premature overconfidence that would amplify annotation noise. The performance drop at $\tau = 0.1$ for 1C1C is particularly pronounced (-2.03 mIoU vs. $\tau = 0.7$), confirming that temperature must be adapted to the annotation budget. In all main experiments, we use the respective optimal temperature for each setting.

B Detailed component design ablations

Table 6 evaluates the three incremental components of FLPE. Starting from the Base model (58.29 mIoU), the frozen LLM prototype module alone provides a +2.24 gain (60.53), directly validating that pretrained text-only attention geometry induces meaningful class prototypes from visual tokens without any LLM weight updates or textual class descriptions.

Adding 2D positional encoding contributes +0.84 (60.53 $\rightarrow$ 61.37): since the LLM is pretrained on sequential text with inherent 1D causal structure, injecting 2D spatial positional embeddings restores spatial adjacency—critical for event data where semantic objects occupy contiguous regions.

Suppressing the causal attention mask adds the final +0.38 (61.37 $\rightarrow$ 61.75). Autoregressive attention restricts each token to preceding positions, essential for text generation but harmful for spatial clustering where full spatial context determines semantic class. The cumulative +3.46 gain confirms the effectiveness of FLPE.

Table 7 investigates FOPA. Starting from LLMBase (61.75 mIoU), adding contrastive loss $\mathcal{L}_{\text{align}}$ without background suppression degrades performance by -0.12 mIoU. This counterintuitive result is the key motivation for Innovation II.

The degradation occurs because ego-motion generates abundant background events that differ substantially between Stream A (long window) and Stream B (short window). When background classes are included as negatives in the InfoNCE loss, the two streams' background prototypes repel each other due to this distribution shift, distorting the prototype geometry on the linguistic manifold and harming foreground separation.

Table 7: Ablation of FOPA components on DDD17 (1C10C).

LLM Base	Contrastive loss	Background ignoring	mIoU (%)
✓			61.75
✓	✓		61.63
✓	✓	✓	**62.55**

Table 8: Independent and joint roles of FLPE and FOPA on DDD17 under 1C10C.

Model variant	mIoU (%)	Δ vs. Base
Dual-stream Base	58.29	–
Single-Stream + FLPE, without FOPA	58.89	+0.60
Dual-Stream + FOPA, without FLPE	60.34	+2.05
Dual-Stream + FLPE + FOPA	**62.55**	**+4.26**

Once background classes are excluded from the contrastive loss, performance jumps to 62.55 mIoU (+0.92 over LLMBase, +1.92 over the contrastive-only variant). This confirms three insights: (1) foreground prototypes benefit from cross-stream mutual information maximization; (2) the presence of background noise due to the nature of the event camera is detrimental as a negative factor in the contrast loss; and (3) mask-driven background exclusion is essential for stable prototype learning under real-world ego-motion.

C Independent Contributions of FLPE and FOPA

Table 8 separates the contributions of semantic prototype formation and temporal prototype alignment under the DDD17 1C10C setting. The Single-Stream+FLPE variant removes the complementary backward stream and therefore cannot exploit cross-stream prototype alignment. Nevertheless, it reaches 58.89 mIoU, exceeding the 58.29 mIoU dual-stream Base by 0.60 points despite using less temporal information. This result indicates that the frozen pretrained layer can improve class structure even when only one event stream is available. Its performance remains well below the complete framework, however, because a single temporal view provides no mechanism for identifying prototype components that are stable across different event accumulations; consequently, the resulting prototypes remain susceptible to stream-specific ego-motion and event-density noise.

Conversely, Dual-Stream+FOPA without FLPE retains both temporal branches but replaces the LLM-derived prototypes with prototypes obtained directly from visual features through masked spatial pooling. This variant reaches 60.34 mIoU, a 2.05-point improvement over the Base, showing that cross-stream foreground alignment and background exclusion are beneficial independently of the LLM prior. However, it remains 2.21 points below the full model because sparsely supervised visual prototypes lack the structured interaction prior supplied by FLPE. Combining both components yields 62.55 mIoU, improving over the Base by 4.26 points. These comparisons reveal complementary functions rather than redundant gains: FLPE supplies class-discriminative structural guidance under sparse supervision, whereas FOPA selects and aligns the foreground structure that remains consistent across temporally different event streams.

D Feature transformation versus prototype formation

We further examine whether the frozen LLM contributes merely by acting as an additional feature-mixing block. For this control, we keep the frozen pretrained LLM layer and the surrounding visual framework unchanged, but remove the learnable class-query tokens. The LLM therefore processes only the projected spatial tokens, and its outputs are returned to the segmentation pathway as transformed visual features; no class-indexed representations are explicitly extracted from the LLM output. This feature-transformer variant achieves 61.48 mIoU, improving over the 58.29 mIoU visual Base by 3.19 points. The gain confirms that pretrained attention is useful even as a generic

Table 9: Feature transformation versus explicit prototype formation on DDD17 under 1C10C.

Model variant	**mIoU (%)**	**Δ vs. Base**
Dual-stream Base	58.29	–
LLM as feature transformer	61.48	+3.19
LLM as prototype regularizer	**62.55**	**+4.26**

interaction operator and that the benefit of the LLM cannot be reduced to the class-query mechanism alone.

Explicit prototype regularization nevertheless raises performance to 62.55 mIoU, providing a further 1.07-point improvement under the same annotation budget. Unlike implicit feature mixing, the proposed design assigns one learnable query to each semantic class. Through joint attention with the spatial tokens, each query aggregates class-relevant evidence into a compact prototype, and the resulting prototype–pixel similarity maps are supervised directly at the annotated locations. This class-indexed bottleneck converts the frozen LLM's interaction structure into an explicit metric space for segmentation, rather than requiring the downstream decoder to recover class organization implicitly from transformed features. The comparison therefore isolates the contribution of prototype formation itself: the frozen layer provides a useful pretrained interaction prior in both variants, while class queries and prototype-level supervision make that prior more directly usable under extremely sparse point labels.