Towards Robust and Interpretable Event–Frame Fusion for Autonomous Driving

Dongyue Lu[1,2]

[1]National University of Singapore
[2]IPAL, CNRS IRL 2955, Singapore
dongyue.lu@u.nus.edu

Abstract

Autonomous driving must handle motion blur, low light, and fast-changing scenes, where RGB frames and event cameras provide complementary strengths. This thesis explores how to fuse them across the perception–reasoning–planning pipeline. It introduces **FlexEvent**, a frequency-robust detector with adaptive fusion and label-efficient training; **Talk2Event**, the first benchmark for event–language grounding with attribute-aware modeling; and the **EventDrive**, an event–frame VLM covering the full driving loop. Together, these contributions advance robust perception, interpretable reasoning, and reliable planning for safety-critical driving through event–frame fusion.

Introduction

Autonomous driving offers significant societal benefits, but its perception stack is strained by high speeds, low-light conditions, and occlusions. RGB cameras capture rich appearance yet suffer temporal sparsity and motion blur, while event cameras provide microsecond motion cues but lack texture and semantics. These complementary properties motivate **event–frame fusion**, whose central challenge is adapting to scene frequency, illumination, and sensor reliability while integrating fused perception into reasoning and planning in an interpretable and effective manner.

We conduct a system-level investigation aligned with the **perception–reasoning–planning** loop. On perception, we introduce **FlexEvent**, a *frequency-aware detector* addressing the prevailing limitation that either downsample event streams to frame rate or rely on static fusion policies that deteriorate at high frequencies. By combining *frequency-adaptive fusion* and *self-training on high-frequency pseudo-labels*, our method sustains accuracy under rapid dynamics while maintaining efficiency. Beyond perception, we advance to language-driven reasoning with **Talk2Event**, which formulates a grounded understanding of dynamic scenes through [1]*appearance,* [2]*motion status,* [3]*egocentric relations, and* [4]*inter-object relations*, and employs an attribute-aware framework to fuse multi-attribute representations for precise temporal and spatial grounding. Motivated by recent progress in vision–language modeling, we further pursue **EventDrive**, a domain-specialized VLM that unifies [1]*perception,* [2]*understanding,* [3]*prediction, and* [4]*planning* via fine-grained, complementary event–frame fusion. Our overarching goal is to establish the value of event sensing in modern driving systems, enabling more robust perception, clearer reasoning, and coherent guidance throughout the driving pipeline.

Related Work

Event-based perception. Event cameras have enabled robust dynamic-scene understanding under high speed or low light, supported by benchmarks such as DSEC (Gehrig et al. 2021) and extended to detection and segmentation methods (Gehrig and Scaramuzza 2023). While these studies advance event-only modeling, they remain limited in semantic richness and evaluation in frame-rate labels, and thus can not fully leverage the high-resolution nature of event cameras.

Event–frame multimodal learning. Combining events with RGB frames has proven effective in tasks such as deblurring and detection. Recent methods explore feature-level fusion with attention or transformers (Gehrig and Scaramuzza 2024), yet struggle with interpretable modality contributions and maintaining efficiency.

Event-based VLMs. Recent works try to align events with language for captioning and QA (Liu et al. 2025; Zhou and Lee 2025), but they do not address spatial grounding or planning-related reasoning. This leaves open opportunities for event-centric VLMs specialized for autonomous driving, where perception, spatial relations, and ego-planning can be integrated under one language-native interface.

Integrating Events into Driving Intelligence

Event cameras provide microsecond dynamics robust to motion blur and low light, while RGB frames offer semantic richness but are temporally sparse; fusing them is promising. Yet current detectors either collapse events to frame-rate supervision or rely on static fusion policies that fail under frequency shifts, and dense high-frequency labels are costly. To address this, we propose **FlexEvent**, a frequency-robust detector that fuses asynchronous events and RGB frames through **FlexFuse**, an *input-adaptive fusion head* with gated soft weights, and **FlexTune**, a *label-efficient adaptation pipeline* with pseudo-labels, temporal calibration, and cyclic self-training. Experiments show real-time

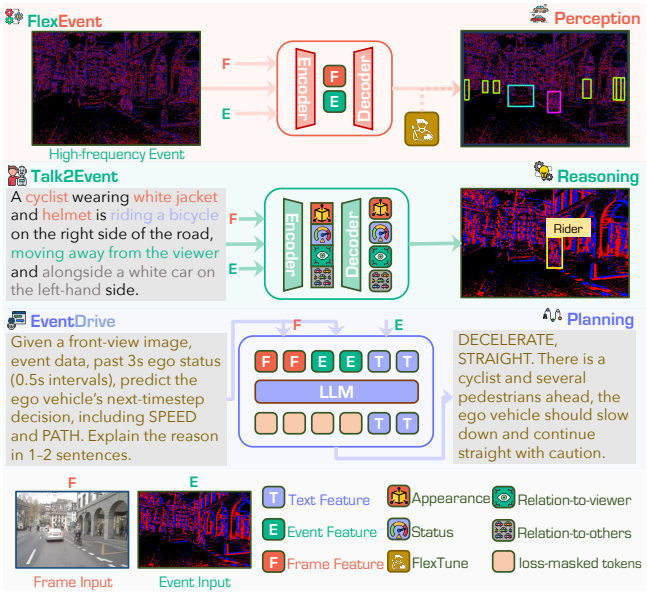

Figure 1: Overview of the proposed research path. **FlexEvent** achieves event–frame fusion through gated soft weights and frequency-adaptive fine-tuning for label-efficient adaptation. **Talk2Event** advances grounded reasoning by aligning events with language via four motion-centric attributes using a mixture-of-attribute experts model. **EventDrive** unifies event, frame, and text features into a shared embedding space, producing language-native outputs that connect perception to reasoning and planning.

performance across wide frequencies with consistent gains over unimodal and fusion baselines. This paper has been accepted to **NeurIPS 2025**, and I am the first author.

Perception alone yields categories and boxes but cannot resolve which instance a query refers to or how motion and relations affect intent. This gap is critical in dynamic, low-light settings where event streams provide rich dynamics but lack semantic grounding. To bridge robust detection with language-driven reasoning, we introduce **Talk2Event**, a large-scale benchmark for **event–language grounding** with referring expressions annotated by motion-centric attributes, including *appearance, motion status, egocentric relation, and inter-object relation*. We further propose **EventRefer**, a **mixture-of-attribute experts** model that fuses event and text representations for precise grounding and interpretable attribution. Experiments establish strong baselines and attribute-level diagnostics. This work is accepted to **NeurIPS 2025 (Highlight)**. I co-lead dataset construction, method formulation, and evaluation.

While perception and grounding improve recognition in dynamic scenes, they do not provide the sequential reasoning or planning required for a complete driving loop. Vision–language models (VLMs) offer a unified interface that connects perception with higher-level interpretation, yet current designs are not tailored for event–frame fusion, where temporal fidelity and motion cues are crucial. To address this gap, we present **EventDrive**, a full-stack event–frame driving framework that unifies [1]*perception*, [2]*understanding*, [3]*prediction*, and [4]*planning* within a language-grounded formulation. EventDrive combines asynchronous events with RGB frames to support core driving tasks: describing scene states, interpreting object semantics and spatial relations, anticipating agent behavior, and suggesting ego maneuvers from trajectories and context. By coupling the temporal precision of events with the semantic richness of frames, the framework enables coherent and robust multimodal reasoning under challenging conditions. This work has been submitted to **CVPR 2026**, with me as the first author.

In the future, we aim to extend **EventDrive** with *iterative perception* guided by event streams, enabling the model to focus on uncertain regions in long-horizon driving. **Reinforcement fine-tuning** will optimize this policy to reward temporal consistency and capture subtle motion cues. Leveraging microsecond event dynamics, this approach promises more human-like perception loops with greater robustness, interpretability, and reliability under challenging conditions. **Timeline.** By **late 2025**, I will release **EventDrive** and the dataset. In the **first half of 2026**, I plan to extend EventDrive with iterative perception and reinforcement learning for long-horizon planning. In the **second half of 2026**, I will generalize the framework beyond autonomous driving to broader embodied and human-centric tasks. **By 2027**, I will integrate my three main contributions: FlexEvent for robust perception, EventDrive for reasoning and planning, and its RL-based iterative extension, into a coherent thesis narrative, and in the **second half of 2027**, I will focus on writing, releasing code and datasets, and defending my dissertation.

Conclusion

This work advances event–frame fusion from perception to reasoning and planning, spanning frequency-adaptive detection, multimodal grounding, and a domain-specialized VLM for driving. Future directions include iterative perception with reinforcement fine-tuning for long-horizon reasoning with dynamic event cues. All together, the thesis will consolidate into a unified framework of event–frame fusion for robust, interpretable, and actionable autonomous driving.

References

Gehrig, D.; and Scaramuzza, D. 2024. Low-latency automotive vision with event cameras. *Nature*.

Gehrig, M.; Aarents, W.; Gehrig, D.; and Scaramuzza, D. 2021. DSEC: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*.

Gehrig, M.; and Scaramuzza, D. 2023. Recurrent vision transformers for object detection with event cameras. In *CVPR*.

Liu, S.; Li, J.; Zhao, G.; Zhang, Y.; Meng, X.; Yu, F. R.; Ji, X.; and Li, M. 2025. Eventgpt: Event stream understanding with multimodal large language models. In *CVPR*.

Zhou, H.; and Lee, G. H. 2025. LLaFEA: Frame-Event Complementary Fusion for Fine-Grained Spatiotemporal Understanding in LMMs. In *ICCV*.