Visual Grounding from Event Cameras

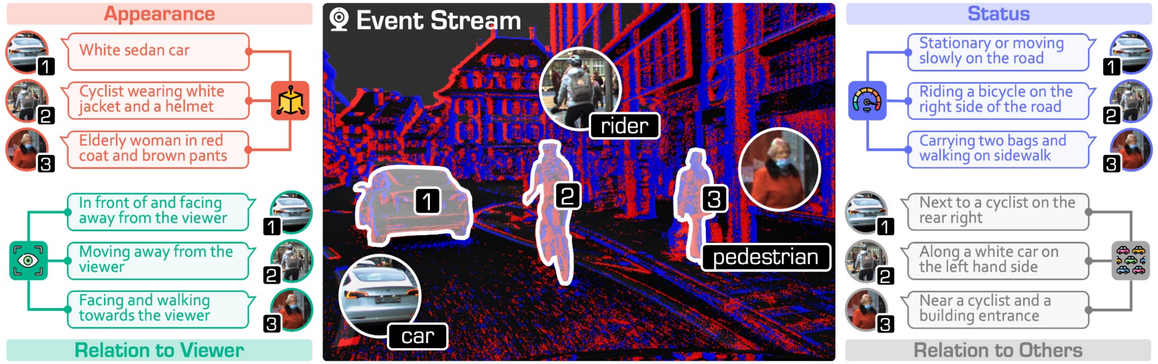
Abstract Paper @ ICCV 2025 NeVi Workshop

Lingdong Kong*, Dongyue Lu*, Ao Liang*,
Rong Li, Yuhao Dong, Tianshuai Hu,
Lai Xing Ng, Wei Tsang Ooi, Benoit R. Cottereau

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Main Motivation & Key Contributions

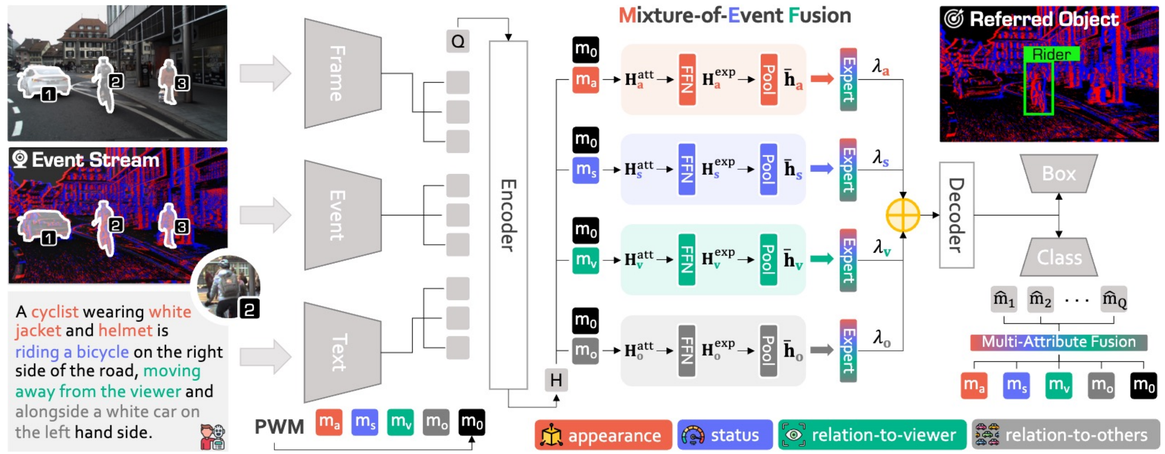
❖ This work introduces **Talk2Event**, a dataset that facilitates language-driven object grounding from event camera data. To our knowledge, this is the **first** benchmark in this kind of research.



❖ The **Talk2Event** dataset consists of **5,567 scenes**, **13,458 annotated objects**, and more than **30,000** carefully validated **referring expressions** for training and testing grounding models.

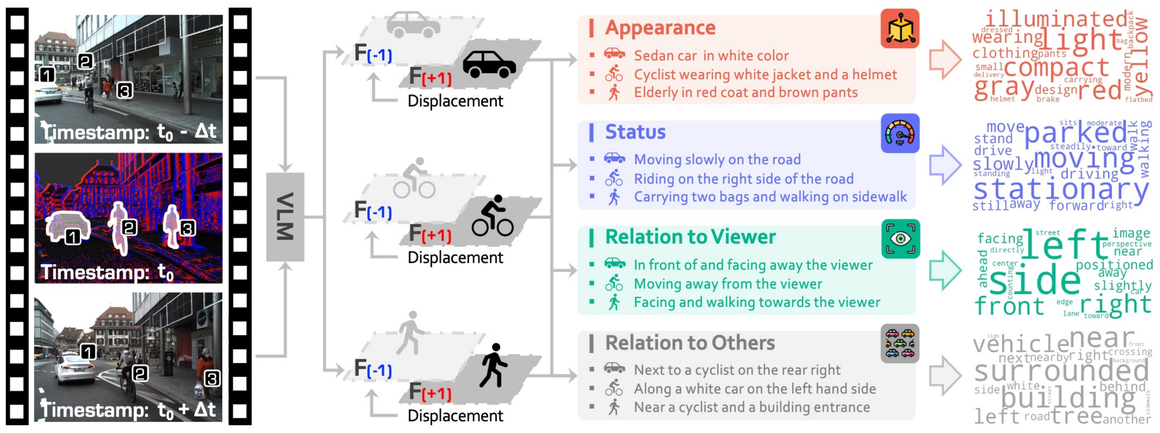
EventRefer: Event-Based Visual Grounding Model

❖ **EventRefer** is an **attribute-aware** grounding model with a mixture of event-attribute experts, achieving state-of-the-art performance across **event-only**, **frame-only**, and **fusion** settings.



Dataset Curation & Annotation Process

❖ Each expression in **Talk2Event** is enriched with 4 structured attributes: [1]**appearance**, [2]**status**, [3]**relation to the viewer**, and [4]**relation to surrounding objects**, that explicitly capture spatial, temporal, and relational cues. This attribute-centric design supports better **interpretable and compositional grounding**, enabling analysis that moves beyond simple object recognition to contextual reasoning in dynamic environments, which is important for event-based vision.

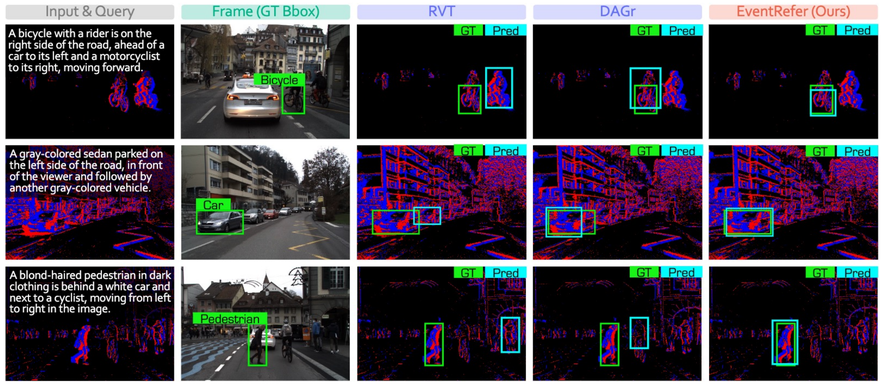


Experiments & Observations

❖ We benchmark 3 groups of methods on **Talk2Event**, including traditional visual grounding methods, **EventRefer,** and zero-shot visual grounding from the large-scale generalist models.

❖ Notably, we observe substantial improvements of our baseline on grounding the **small or dynamic objects**, such as **pedestrians (+5.0%)** and **riders (+24.4%)**, demonstrating the ability to leverage attribute-level reasoning beyond the simple appearance-based matching.

❖ Grounding motion dynamics can bring unique advantages, especially for **low-light** or **high-speed** scenes, where the frame-based model tends to struggle.



❖ We hope this work can open up more possibilities for the development of more advanced event-based scene understanding systems.

DesCartes · cnrs @CREATE Singapore · NUS National University of Singapore · Nanyang Technological University Singapore · Agency for Science, Technology and Research Singapore · 香港科技大學 The Hong Kong University of Science and Technology