Segment Any Events with Language

Seungjun Lee, **Gim Hee Lee**

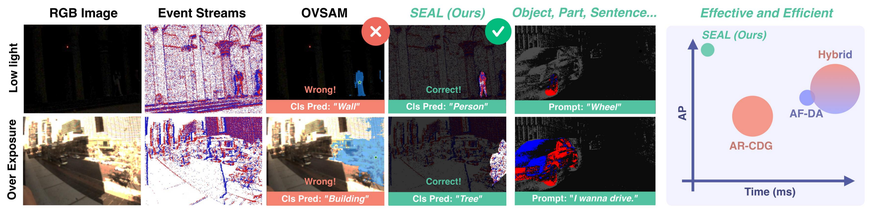
Department of Computer Science, National University of Singapore

ICLR 2026

Project Page

Linkedin

Introduce our SEAL



(a) RGB vs Event

(b) SEAL (c) Comparison

RGB camera alone is not enough in the real-world!

- Even though we are living in the era of 2D / 3D foundation models, their robustness in real-world conditions remains an open question.
- They often make false predictions in the challenging conditions such as low-light, motion blur, or overexposure.

Event cameras can be the solution!

Events naturally offer high dynamic range and temporal resolution which makes it robust to challenging conditions. → **Events sensor can mitigate errors of RGB models!**

However, understanding events with open-vocabulary or natural language remains largely underexplored.

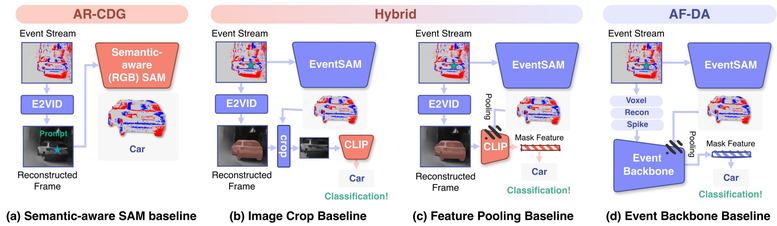
The first, Open-Vocabulary Segment Any Events

- Given visual prompt such as point or box, our SEAL performs segmentation and mask classification with open-vocabulary queries.
- Can understand multiple levels of granularity: both part-level and instance-level
- Real-time inference!
- Parameter-efficient architecture!

New benchmarks

Proposes four benchmarks to evaluate our SEAL in diverse settings:
1) Label granularity: 6classes → 11 classes → 19 classes
2) Semantic granularity: Instance → Part

Baseline

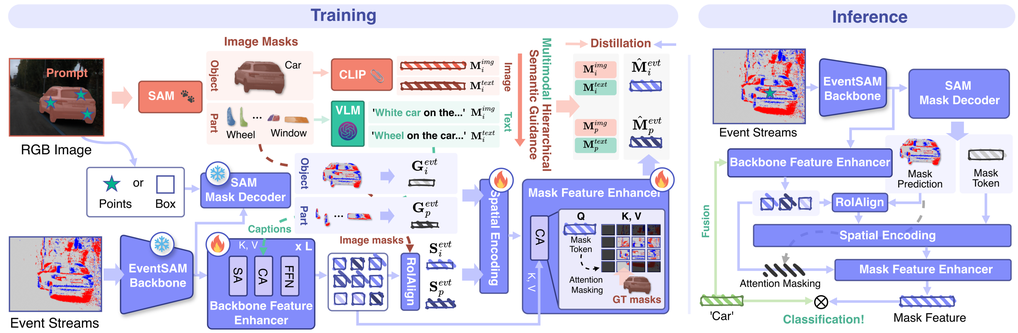


(a) Semantic-aware SAM baseline (b) Image Crop Baseline (c) Feature Pooling Baseline (d) Event Backbone Baseline

Limitation

- Too slow. Not real-time.
- Too heavy.
- Suffer from domain gap.

SEAL (Ours)



Backbone Feature Enhancer

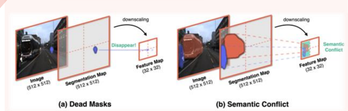
Injecting language prior into events backbone via CA.

→ Better alignment between events and language

Spatial Encoder

Only using enhanced backbone features are not enough.

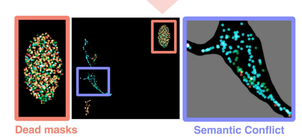
1) Dead masks problem
2) Semantic conflict problem



→ Add geometric prior from SAM's mask decoder to solve those issues.

Mask Feature Enhancer

Further enhance spatial and semantic prior via masked CA layer.



SE and MFE layer lead to more distinct events feature space

MHSG Guidance

Multimodal learning framework to learn semantic-rich event representations.

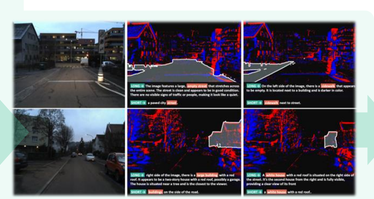
Multimodal **Multi-granularity**

Using multiple foundation models:

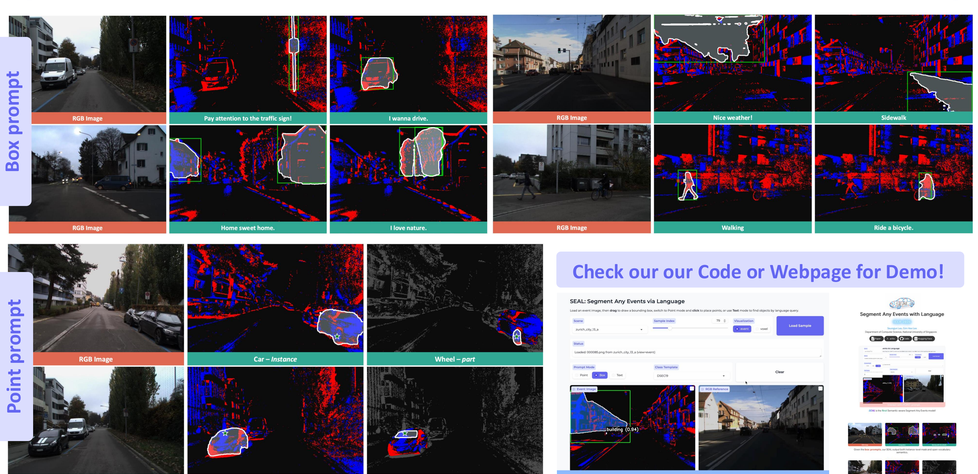
OpenSeg → Visual guidance

Osprey → Text guidance

SAM → *Semantic*, *Instance*, *Part*-level masks

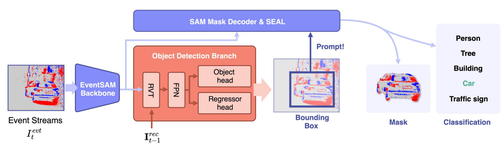


Interact with events



Check our our Code or Webpage for Demo!

SEAL++



We further introduce SEAL++!

- SEAL works in one events frame (Events image). SEAL++ works in events video!
- Spatiotemporal, prompt-free OV-EIS.
- Check our webpage to see the qualitative video!

Class-agnostic object detection *I wanna drive.* *Walking in the sidewalk.*