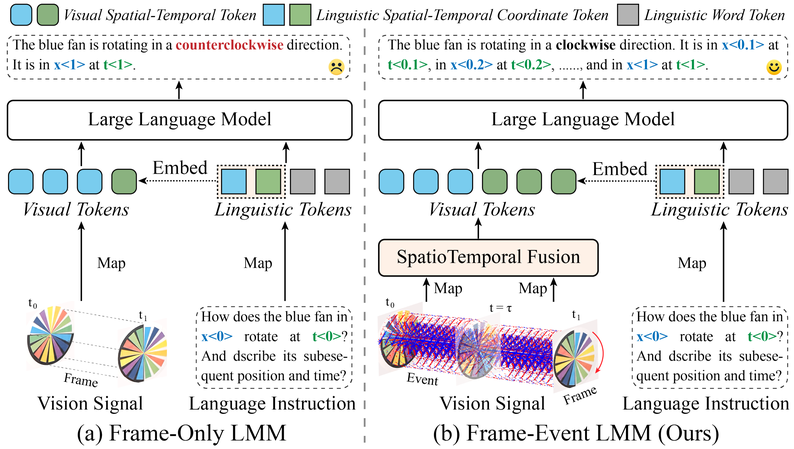
LLaFEA: Frame-Event Complementary Fusion for Fine-Grained Spatiotemporal Understanding in LMMs

Hanyu Zhou, Gim Hee Lee*
National University of Singapore

ICCV OCT 19-23, 2025 HONOLULU HAWAII

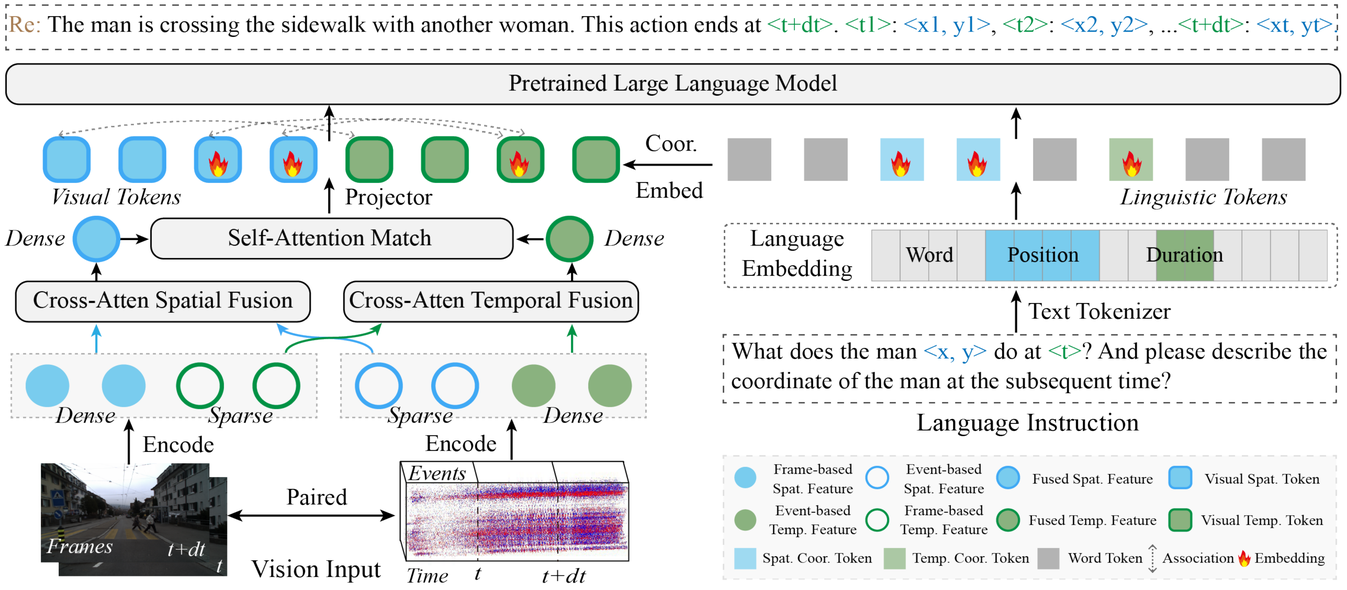
Problem & Idea

Frame-only LMMs suffer from temporal sparsity that weakens vision-language alignment. We incorporate a temporally dense event camera to enhance spatiotemporal representation for LMMs to interpret scenes accurately at any position and time.



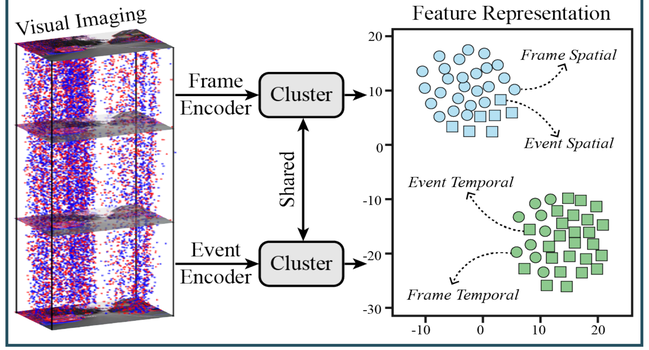
Large Language and Frame-Event Assistant (LLaFEA)

- **Visual Frame-Event Spatio-temporal Fusion:** Cross-attention fusion and self-attention match for visual features.
- **Language-Vision Coordinate Align:** Embedding spatio-temporal coordiantes into the fused visual features.

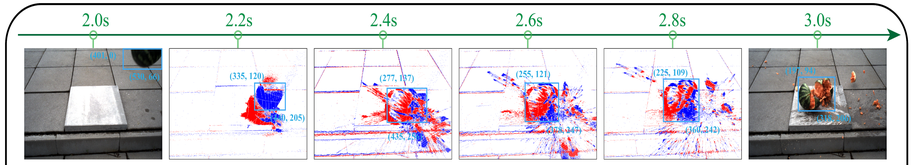


Motivation

Frames capture a **spatially dense** and **temporally sparse** global appearance. Events detect **spatially sparse** and **temporally dense** local boundaries.



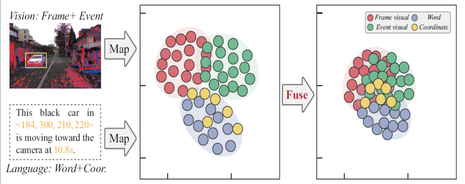
Comparison on Scene Understanding



User	Describe the position varying of the dynamic object during <2.0s-3.0s>. What may happen at 2.4s?
GroundingGPT	This melon in <399,5, 528, 59> starts falling at 2.0s, and shatters on the ground in <189, 90, 325, 224> at 3.0s. At 2.4s, it is hurtling toward the ground with rapid acceleration.
GroundingGPT w/ FI	This melon in <399,5, 528, 59> starts falling at 2.0s, and is in <291, 151, 410, 236> at 2.2s, ..., and breaks on the ground in <187, 88, 319, 226> at 3.0s. At 2.4s, it has broken into pieces on the ground.
LLaFEA	The dynamic object is a **green watermelon**. This watermelon is beginning to fall in <400, 1, 528, 63> at 2.0s, and is about to hit the ground in <334, 124, 435, 201> at 2.2s, ..., and has crashed to the ground with small pieces flying around in <195, 94, 320, 201> at 3.0s. At 2.4s, **this watermelon is disintegrating in** <274, 134, 432, 258>, its structure breaking apart due to the overwhelming force of the impact.

Ablation Study and Discussion

Method		Video-ChatGPT	Video-LLaVA		GroundingGPT		Grounded-VideoLLM		VTimeLLM		Qwen2.5-VL	InternVL 2.5	LLaVA-OneVision		Ours
		–	–	w/ FI	–	w/ FI	–	w/ FI	–	w/ FI	w/ FI	w/ FI	w/ FI		
Vision input		F	F	F+E	F	F+E	F	F+E	F	F+E	F+E	F+E	F+E		F+E
Syn Event -ST	DC↑	2.51	2.14	2.75	2.58	2.91	2.64	3.27	2.83	3.30	3.55	3.37	3.42		**3.94**
	VQA↑	2.80	2.56	3.12	2.85	3.24	2.96	3.63	3.10	3.84	4.19	4.21	4.05		**4.72**
	s-IoU↑	–	–	–	4.23	9.70	–	–	–	–	–	16.25	12.37		**56.82**
	t-IoU↑	–	–	–	6.94	11.54	7.20	12.15	34.61	39.09	–	–	–		**68.35**
Real Event -ST	DC↑	1.98	1.82	2.30	2.26	2.55	2.20	2.80	2.45	2.92	3.14	2.98	2.85		**3.65**
	VQA↑	2.31	1.80	2.45	3.04	3.62	3.11	3.75	3.18	3.93	4.05	4.01	3.87		**4.82**
	s-IoU↑	–	–	–	4.02	7.97	–	–	–	–	–	13.29	9.30		**46.27**
	t-IoU↑	–	–	–	6.23	10.61	5.35	7.40	21.05	23.55	–	–	–		**52.80**



Cross-Attention		Self-Attention Matching	DC↑	VQA↑	s-IoU↑	t-IoU↑
Spatial	Temporal					
×	×	×	2.25	3.10	4.03	6.15
√	×	×	2.39	3.27	15.40	6.16
×	√	×	2.30	3.25	4.53	22.74
×	×	√	3.08	4.10	12.35	21.02
√	√	√	**3.65**	**4.82**	**46.27**	**52.80**