SPARSE FRAMES, CAUSAL GRAPHS: LONG EVENT-STREAM ANOMALY REASONING WITH VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

RGB sensing can lose critical anomaly evidence under fast motion or adverse illumination, while event cameras preserve high-dynamic-range, microsecond-scale changes. Most event-language methods nevertheless aggregate events into frame-like inputs, obscuring the temporal detail that makes event sensing valuable. We propose EFG-VLM, an offline hybrid event-frame/event-graph framework for long-stream open-world anomaly understanding. A category-agnostic locator establishes shared anchors where sparse event frames provide scene semantics and bounded raw-event graphs preserve complementary spatial structure and causal motion. Hierarchical fusion combines landmark, proposal, and video context and exposes structured evidence to a pretrained multimodal language model for anomaly judgement, proposal-conditioned temporal localization, and description. Because localization is category-agnostic and final interpretation is language-generated rather than closed-set classification, the same interface can describe interactions outside benchmark categories. On large-scale v2e-simulated event streams derived from HIVAU-70K, EFG-VLM substantially improves event-source recognition and language description while remaining robust across clip durations. Native RGB–event cases further illustrate transfer to high-dynamic-range interactions beyond the benchmark categories.

1 INTRODUCTION

Video Anomaly Detection (VAD) requires both locating unusual events and interpreting them under scene-dependent semantic and physical context. Recent MLLM-based methods improve open-world reasoning through visual-language priors Tang et al. (2024); Zhang et al. (2025); Liu et al. (2026b), but still rely largely on RGB video. Under low light, exposure changes, or rapid motion, RGB sensing may blur or miss short-lived interactions such as impacts and collisions Gallego et al. (2022); Qian et al. (2025); once this evidence is lost, stronger semantic reasoning cannot reconstruct the missing temporal detail.

Event cameras complement RGB sensing with high dynamic range and microsecond-level temporal resolution, directly preserving rapid brightness changes that conventional video may miss. Yet their sparse, irregular streams are poorly matched to frame-oriented MLLMs, so existing Event-MLLMs commonly aggregate events into frame-like representations Liu et al. (2025b); Cao et al. (2026), gaining semantic compatibility at the cost of temporal quantization.

This motivates a complementary design that separates semantic anchoring from event-native motion modeling. Semantic reasoning benefits from sparse frame-like observations that provide stable scene and object references compatible with pretrained visual encoders, whereas attributing short physical anomalies requires preserving the fine temporal ordering and local interactions of asynchronous events. We therefore use event-frame anchors for semantic context and spatial–motion event graphs for fine-grained physical evidence, and fuse the two only after their complementary information has been preserved. In this work, we focus on *offline long-stream anomaly understanding* rather than strictly causal early-warning detection. Given a recorded event stream, our goal is to determine what abnormal interaction occurs, when it occurs within the stream, and how it should be interpreted under its semantic context. This setting allows the temporal locator to use a bounded amount of past

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.

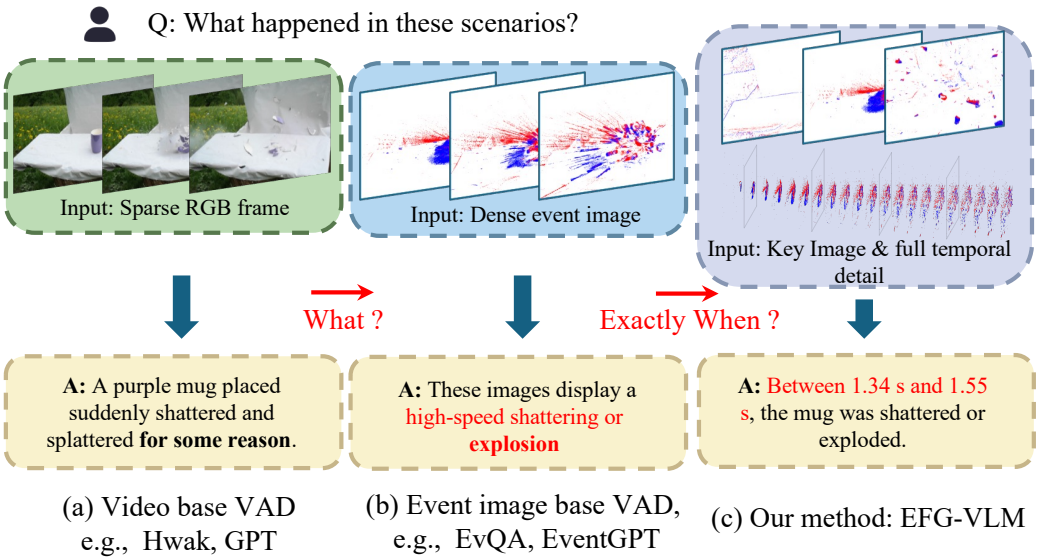


Figure 1: **Motivation and comparison of anomaly understanding from different visual representations.** RGB-based VAD can miss short, high-speed changes, while event-image-based methods preserve motion cues but compress fine temporal structure. EFG-VLM combines sparse semantic anchors with the full temporal detail of the event stream to identify both *what* happened and *when* it occurred.

and future context for proposal localization, while the event representation within each selected local window remains explicitly causal in its motion modeling.

We propose EFG-VLM, a hybrid event-frame and event-graph framework for long-stream open-world anomaly understanding. A class-agnostic temporal locator first identifies candidate activities and establishes five shared temporal anchors. Around the same anchors, sparse event frames provide scene semantics, while raw-event graphs separately model near-simultaneous spatial structure and directed motion. The resulting landmark evidence is fused with longer-range temporal context and exposed to Qwen through structured proposal, landmark, prior, and video-context tokens.

For evaluation, we adopt HIVAU-70K from Holmes-VAU (Zhang et al., 2025), convert its source RGB videos to event streams with v2e (Hu et al., 2021), and report results separately on the XD-Violence and UCF-Crime source subsets. We further evaluate long-input robustness on a duration-stratified HIVAU-70K subset and open-world transfer on native synchronized RGB–event sequences.

In summary, our contributions are threefold:

- We introduce EFG-VLM, which combines sparse event-frame semantics with event-native spatial–motion graphs for offline long-stream anomaly understanding.
- We develop a proposal-guided hierarchical interface that aligns both branches with shared temporal anchors, integrates local motion evidence with proposal-level and video-level context, and conditions temporal localization on structured proposal priors.
- We reveal complementary roles of spatial structure and causal motion topology in event-language grounding: structural evidence supports anomaly recognition, while directed motion evolution substantially improves semantic description, yielding strong gains over existing event-source baselines and successful transfer to native high-dynamic-range scenes.

2 Related Work

Video Anomaly Detection. Classical VAD methods learn visual or temporal regularities from RGB video Sultani et al. (2018); Hirschorn & Avidan (2023); Cao et al. (2024), while recent MLLM-based methods add language-grounded open-world reasoning Tang et al. (2024); Zhang et al. (2025); Yang et al. (2026); Liu et al. (2026b); Cai et al. (2026). Representative systems emphasize motion cues, hierarchical temporal annotation, or anomaly-sensitive attention, but remain dependent on RGB evidence that can degrade under adverse illumination or fast motion. Our work instead preserves low-level temporal evidence before language reasoning.

Event-based Visual Representation. Most event-vision pipelines convert asynchronous streams into event frames, time surfaces, or voxel grids before applying CNN or ViT encoders Liang et al. (2024); Liu et al. (2025b); Zhou & Lee (2025); Lou et al. (2025); Xu et al. (2025); Liu et al. (2026a); Cao et al. (2026). Such representations are efficient and compatible with pretrained image models, but temporal aggregation inevitably introduces quantization. Graph-based methods instead preserve local event geometry and motion through spatiotemporal relations and have been explored for recognition, optical flow, and dynamic prediction Deng et al. (2022); Schaefer et al. (2022); Dalgaty et al. (2023); Dampfhoffer et al. (2025). Because a graph over a complete long stream is prohibitively expensive, EFG-VLM restricts event-native graph reasoning to proposal-guided windows while retaining a coarse timeline for long-range context.

Events with Foundation Models. EventGPT, EventFlash, EvQA, and EventBench extend event sensing toward captioning, VQA, and general event-language understanding Liu et al. (2025b; 2026a); Lou et al. (2025); Liu et al. (2025a), using multi-stage alignment, adaptive aggregation, or reconstruction-based bridging. Their success shows that events can align with foundation-model semantics, but most interfaces still depend on reconstructed or temporally aggregated frame-like inputs. RGB-event systems such as LLaFEA (Zhou & Lee, 2025) and EventSTU (Xu et al., 2025) further show that event dynamics can complement conventional visual semantics. In contrast, EFG-VLM targets long-stream anomaly understanding using only the event stream at inference and explicitly separates sparse semantic anchors from event-native graph dynamics before hierarchical fusion.

3 Methodology

EFG-VLM balances a practical tension in long event streams: full-stream graphs are expensive, while frame aggregation can erase short dynamics. We therefore localize candidate intervals, establish shared anchors for sparse semantic frames and bounded raw-event graphs, and fuse both with longer-range context for anomaly judgement, proposal-conditioned temporal localization, and description. The four-stage pipeline comprises proposal-guided anchors, spatial–motion graph encoding, hierarchical context encoding, and language grounding (Fig. 2); low-level details are in Appendix A.

3.1 Proposal-Guided Shared Temporal Anchors

Given an event stream $\mathcal{S} = \{e_i\}_{i=1}^{N}$ with $e_i = (x_i, y_i, t_i, p_i)$, we first construct a 4-Hz coarse timeline from 0.25-second event bins. A temporal convolutional encoder with symmetric dilated convolutions predicts eventness, temporal boundaries, centers, and offsets, from which candidate intervals are refined and filtered into

$$\mathcal{P} = \{[s_p, e_p]\}_{p=1}^{P}. \tag{1}$$

The locator is category-agnostic and uses bounded bidirectional timeline features; proposal generation and ranking over the folded timeline are therefore offline rather than strictly causal. This is separated from the subsequent event representation, whose motion edges within each selected window remain directed from earlier to later events. The 4-Hz timeline is reused as long-range context, while fine scene and motion evidence is sampled from the original event stream.

A proposal should preserve both its global temporal extent and the short internal transitions that may determine anomaly semantics. We therefore use five shared anchors: start, center $c_p = (s_p + e_p)/2$, and end provide a stable structural scaffold, while two change-aware landmarks capture the strongest

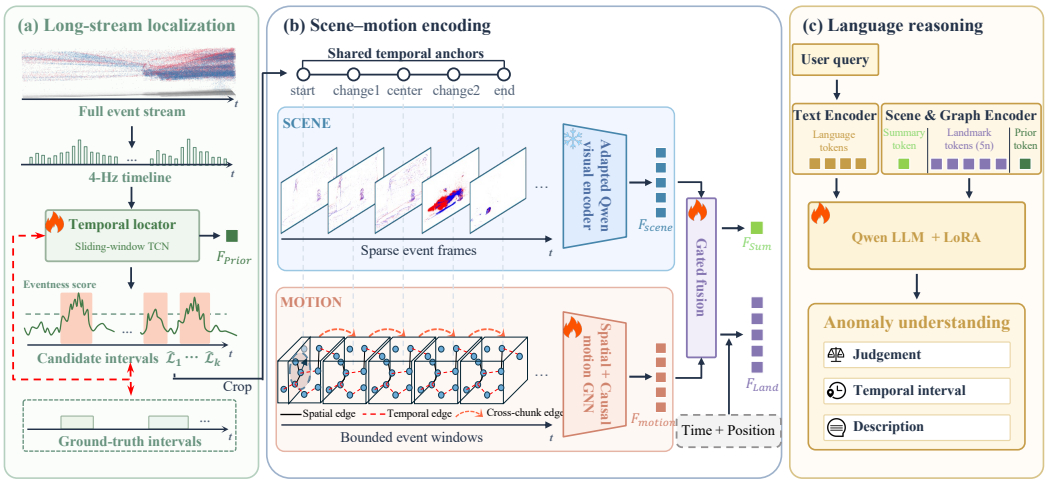


Figure 2: Overview of EFG-VLM. A temporal locator establishes candidate proposals and shared anchors for sparse event-frame semantics and bounded spatial–motion graphs. Proposal-level temporal priors and hierarchical scene–motion evidence are then grounded into Qwen for anomaly judgement, proposal-conditioned temporal localization, and description.

internal variation in the first and second halves of the proposal, respectively. This half-wise design encourages coverage of distinct activity stages while maintaining a compact fixed-budget interface shared by the event-frame and graph branches. With timeline features $\mathbf{h}_m$ at time τ_m, we define $d_m = \|\mathbf{h}_m - \mathbf{h}_{m-1}\|_2$ and select the strongest change in each half of the proposal:

$$t_p^{(1)} = \tau_{\arg\max_{s_p < \tau_m < c_p} d_m}, \qquad t_p^{(2)} = \tau_{\arg\max_{c_p < \tau_m < e_p} d_m}. \tag{2}$$

The ordered anchors are therefore

$$\mathcal{T}_p = (s_p, t_p^{(1)}, c_p, t_p^{(2)}, e_p). \tag{3}$$

At each anchor, a short event window is rendered into an event frame for the adapted Qwen visual encoder, while the same temporal reference determines the bounded raw-event window used by the graph branch. This shared selection keeps scene semantics and event-native dynamics aligned to the same activity. As shown in Table 3, replacing the two change-aware landmarks with uniformly spaced anchors substantially degrades language quality under the same five-anchor budget.

3.2 Spatial–Motion Graph Encoding

Because event frames partially collapse local ordering and burst structure, we construct graphs directly from bounded raw-event windows and retain individual events as nodes up to a fixed budget. This preserves event-level structure without the cost of a graph over the entire video. Each retained node is represented as

$$\mathbf{x}_i = [\hat{x}_i, \hat{y}_i, \bar{t}_i, p_i], \tag{4}$$

where spatial coordinates and chunk-local time are normalized.

Spatial structure and motion impose different relational geometries: near-simultaneous events depend mainly on image-plane proximity, whereas motion reasoning depends on temporal order and elapsed time. A single (x, y, t) distance can mix structural neighbors with motion predecessors, so we construct two relations on the same nodes:

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}^{\text{spa}} \cup \mathcal{E}^{\text{mot}}). \tag{5}$$

Spatial edges connect nearby, near-simultaneous events within the same local chunk, preserving contours and short-term spatial structure. Causal motion edges instead connect earlier events to later targets under bounded temporal and spatial gaps,

$$(i, j) \in \mathcal{E}^{\text{mot}} \Rightarrow 0 < t_j - t_i \le \tau_m, \qquad \|\mathbf{r}_j - \mathbf{r}_i\|_2 \le R_m, \tag{6}$$

and may cross chunk boundaries to preserve short-range motion continuity Dalgaty et al. (2023); Dampfhoffer et al. (2025).

The two relations are processed by separate graph blocks,

$$\mathbf{H}^s = \Phi_s(\mathbf{H}, \mathcal{E}^{\text{spa}}), \qquad \mathbf{H}^m = \Phi_m(\mathbf{H}, \mathcal{E}^{\text{mot}}), \tag{7}$$

and fused by a node-wise gate

$$\mathbf{H}' = \mathbf{H}^s + \sigma(\phi([\mathbf{H}^s; \mathbf{H}^m])) \odot \mathbf{H}^m. \tag{8}$$

After relation-specific propagation, node features are pooled within each chunk using complementary statistics and gated attention. The resulting ordered chunk sequence is processed by a causal multi-scale temporal encoder and a short GRU whose hidden state is reset for every bounded window, so no recurrent state leaks across proposals or videos. A final salient pooling produces the landmark-level motion representation $\mathbf{m}_{p,l}$. Exact neighborhood construction, edge attributes, pooling operators, and sampling budgets are deferred to Appendix A.2.

3.3 Hierarchical Scene–Motion Context Encoding

Because the graph and event-frame branches share the same temporal anchors, their features can be fused without additional temporal matching. For landmark l of proposal p, we use graph-primary residual fusion:

$$\mathbf{z}_{p,l} = \mathbf{m}_{p,l} + \mathbf{g}_{p,l} \odot \tilde{\mathbf{v}}_{p,l}, \qquad \mathbf{g}_{p,l} = \sigma(\phi_g[\mathbf{m}_{p,l}; \tilde{\mathbf{v}}_{p,l}])\,. \tag{9}$$

Here $\tilde{\mathbf{v}}_{p,l}$ is the projected event-frame feature, so scene semantics complement rather than replace event-native motion evidence.

Each landmark also receives a compact temporal coordinate encoding

$$\boldsymbol{\xi}_{p,l} = \left[\frac{t_{p,l} - s_p}{e_p - s_p}, \frac{t_{p,l}}{T}, \frac{e_p - s_p}{T}\right]. \tag{10}$$

A lightweight Transformer processes the ordered five-landmark sequence and yields contextual landmark tokens and a local proposal summary. The explicit ordering preserves the distinct temporal roles of start, change$_1$, center, change$_2$, and end rather than treating the five observations as an unordered set. This summary is combined with coarse before/inside/after context from the 4-Hz timeline and proposal geometry to form $\mathbf{E}_p$.

Because the same local motion can have different meanings under different surrounding events, the proposal summary is interpreted with broader video context obtained by pooling the complete 4-Hz timeline and sparse landmark visuals. A proposal-set Transformer then jointly reasons over this global video token and all retained proposals:

$$[\mathbf{C}^*, \mathbf{E}_1^*, \ldots, \mathbf{E}_P^*] = \text{Transformer}([\mathbf{C}, \mathbf{E}_1, \ldots, \mathbf{E}_P]). \tag{11}$$

Lightweight heads estimate proposal validity, boundary refinement, description readiness, and a perceptual anomaly prior, which serve as structured evidence rather than the final semantic judgement.

3.4 Hierarchical Language Grounding

A single proposal token would discard the internal temporal structure recovered above. We therefore expose, for each proposal, its contextual summary $\mathbf{E}_p^*$, five ordered landmark tokens $\{\mathbf{L}_{p,l}\}_{l=1}^5$, and a numerical temporal prior, together with the global video context $\mathbf{C}^*$. The resulting proposal-major sequence is

$$\mathcal{X} = [\mathbf{C}^*, \mathbf{S}_1, \mathbf{L}_{1,1:5}, \mathbf{P}_1, \ldots, \mathbf{S}_P, \mathbf{L}_{P,1:5}, \mathbf{P}_P], \tag{12}$$

where dedicated projections and type/slot embeddings preserve evidence type and proposal identity. The prior token $\mathbf{P}_p$ carries Stage-1 locator traces plus proposal boundary/confidence cues, so Qwen receives explicit localization evidence rather than only a decoding range.

Qwen combines these structured tokens with the user query and generates an anomaly judgement, a description, and proposal-conditioned temporal localization. Rather than recovering an interval directly from the complete long stream, Qwen predicts normalized start and end bins $\hat{b}_p^s, \hat{b}_p^e \in [0, 100]$ relative to the Stage-1 proposal $[s_p, e_p]$. The predicted bins are mapped back to video time by

$$\hat{s}_p = s_p + \frac{\hat{b}_p^s}{100}(e_p - s_p), \qquad \hat{e}_p = s_p + \frac{\hat{b}_p^e}{100}(e_p - s_p). \tag{13}$$

Table 1: Overall comparison on the XD-Violence and UCF-Crime source subsets of HIVAU-70K. Methods are grouped by source modality to distinguish event-source evaluation from native RGB-video evaluation.

Method	Input	XD-Violence		UCF-Crime		Language quality			
		Acc.	Macro-F1	Acc.	Macro-F1	BLEU	CIDEr	METEOR	ROUGE-L
Event-source input									
EvQA-ART (Lou et al., 2025)	Event recon.	15.18	13.56	10.39	2.04	0.014	0.003	0.172	0.153
EvQA-FRT (Lou et al., 2025)	Event recon.	28.89	30.18	16.91	9.48	0.026	0.005	0.217	0.181
EFG-VLM (ours)	Event stream	50.38	31.14	35.61	8.83	0.064	0.150	0.207	0.239
Native RGB-video input									
HAWK (Tang et al., 2024)	RGB video	24.43	32.68	14.84	14.36	0.029	0.025	0.205	0.187
Holmes-VAU (Zhang et al., 2025)	RGB video	62.79	59.95	60.53	25.39	0.167	0.529	0.367	0.362
MoniTor (Yang et al., 2026)	RGB video	69.42	62.69	90.80	82.78	0.009	0.032	0.101	0.139
LaGoVAD (Liu et al., 2026b)	RGB video	73.39	59.17	46.60	15.57	–	–	–	–
Gemma-4-12B (offline)	RGB video	66.49	62.83	80.71	67.32	0.030	0.101	0.162	0.184
GPT-5.5 (online)	RGB video	83.62	76.44	96.14	95.92	0.021	0.117	0.166	0.188

The pretrained Qwen backbone remains frozen, while the multimodal bridge and LoRA attention adapters are optimized for language grounding. The exact 54-dimensional prior composition and token construction are given in Appendix A.3.

3.5 Multi-Stage Training

Training follows the same perception-to-language decomposition: Stage 1 trains the locator, Stage 2 fixes it and optimizes scene–motion context objectives, and Stage 3 freezes perception while adapting the multimodal bridge and Qwen LoRA modules. Stage 2 representations are cached for instruction tuning; optimization details are in Appendix A.4.

4 Experiments

We evaluate EFG-VLM from four complementary perspectives: overall anomaly recognition and description, robustness to clip duration, timestamped qualitative analysis of semantic interpretation and temporal localization on simulated and native event sequences, and component ablations. Because event-source and RGB-video methods operate on different sensing modalities, we report them in separate groups and do not interpret cross-group score differences as modality-controlled comparisons.

4.1 Experimental Setup

Datasets and tasks. Our primary dataset is HIVAU-70K from Holmes-VAU (Zhang et al., 2025), whose XD-Violence and UCF-Crime source videos are converted one-to-one into event streams using v2e (Hu et al., 2021) while preserving the original annotations and train/test split. All source videos use one fixed v2e configuration; the merged training and test splits contain 160,509 and 33,518 instructions, respectively, with complete simulation parameters given in Appendix B.

Metrics and baselines. We report accuracy and Macro-F1 for anomaly recognition and standard BLEU, CIDEr, METEOR, and ROUGE-L for generated anomaly descriptions. EventGPT (Liu et al., 2025b) is reported only for the shortest-duration group because its supported event segment is approximately 100 ms, while LaGoVAD (Liu et al., 2026b) performs classification only and therefore has no language metrics. GPT-5.5 and MoniTor (Yang et al., 2026) are online API references; refusals/unparsable outputs count as incorrect, and their undisclosed training corpora preclude ruling out prior benchmark exposure. Prompting, simulation, training, and checkpoint details are in Appendix B.

4.2 Overall Comparison

Table 1 shows strong gains over event-source baselines: relative to EvQA-FRT, EFG-VLM raises XD/UCF accuracy from 28.89/16.91 to 50.38/35.61, CIDEr from 0.005 to 0.150, and ROUGE-L

Table 2: Clip-duration-binned robustness on MiniDataset, a duration-stratified subset of HIVAU-70K. Macro-F1 measures anomaly-detection stability and CIDEr measures description stability as clip duration increases. Methods are grouped by source modality. A dash indicates that the corresponding result is unavailable or unsupported.

Method	Input	< 1s		1–5s		5–10s		10–20s		> 20s	
		F1	CIDEr	F1	CIDEr	F1	CIDEr	F1	CIDEr	F1	CIDEr
Event-source input											
EventGPT (Liu et al., 2025b)	Event frames	48.00	0.077	–	–	–	–	–	–	–	–
EvQA-ART (Lou et al., 2025)	Event recon.	7.69	0.000	18.41	0.000	17.06	0.000	17.38	0.002	31.47	0.002
EvQA-FRT (Lou et al., 2025)	Event recon.	53.33	0.000	39.45	0.000	53.54	0.000	46.34	0.002	59.46	0.003
EFG-VLM (ours)	Event stream	50.14	0.129	50.89	0.093	52.51	0.106	64.71	0.085	73.55	0.064
Native RGB-video input											
HAWK (Tang et al., 2024)	RGB video	35.13	0.014	33.84	0.048	36.43	0.034	36.71	0.025	31.69	0.011
Holmes-VAU (Zhang et al., 2025)	RGB video	48.00	0.568	48.23	0.311	73.41	0.318	77.24	0.311	78.87	0.395
MoniTor (Yang et al., 2026)	RGB video	72.92	0.138	74.64	0.084	85.19	0.035	81.72	0.019	87.50	0.005
LaGoVAD (Liu et al., 2026b)	RGB video	48.12	–	56.31	–	76.12	–	86.24	–	90.78	–
Gemma-4-12B (offline)	RGB video	52.97	0.164	67.06	0.124	67.54	0.102	80.06	0.094	85.47	0.061
GPT-5.5 (online)	RGB video	46.94	0.205	75.16	0.141	86.82	0.135	91.35	0.100	95.03	0.062

from 0.181 to 0.239. The trend is less uniform for class-wise recognition: Macro-F1 is 31.14 vs. 30.18 on XD but 8.83 vs. 9.48 on UCF, and METEOR remains slightly lower (0.207 vs. 0.217). Thus, the strongest gains concern overall recognition and semantic description rather than every class or language metric.

RGB-video systems are modality-specific references rather than controlled comparisons: several achieve stronger classification, and Holmes-VAU has higher aggregate language scores. This does not imply that RGB is uniformly superior under the sensing conditions motivating event cameras. In rapid *accident*, the frame-based descriptions capture the crash outcome but leave the two-vehicle interaction unresolved, whereas EFG-VLM recovers the collision itself; in low-light *jenga*, it identifies the incoming object and impact while the compared frame-based descriptions mainly capture the subsequent collapse (Fig. 3). These cases are qualitative rather than modality-controlled quantitative evidence, but show how stronger aggregate language scores can coexist with failures when decisive evidence is brief or visually degraded.

4.3 Performance Across Clip Durations

Table 2 shows stable recognition across duration bins: EFG-VLM Macro-F1 rises from 50.14 in < 1s clips to 73.55 beyond 20s. Description is more sensitive to length, with CIDEr decreasing from 0.129 to 0.064, yet remaining well above EvQA-FRT at > 20s (0.064 vs. 0.003). RGB references show a related recognition/description separation: GPT-5.5 reaches 95.03 Macro-F1 at > 20s while CIDEr is 0.062, and MoniTor's CIDEr falls from 0.138 to 0.005 across the duration range.

Because bins contain different clips, these results compare robustness regimes rather than establish a causal effect of duration; the < 1s bin is also smaller (78 vs. 372 samples elsewhere). EvQA-ART remains below FRT across all bins rather than failing only for < 1s clips. ART sparsely reconstructs regions only after sufficient event activity, whereas FRT produces dense frames at a fixed rate; the former can therefore be more sensitive to event-density statistics. Such statistics may shift under v2e simulation, plausibly amplifying the ART–FRT gap, although we treat this as a method-specific hypothesis because trigger statistics are not measured. EventGPT's missing long-duration entries reflect its input-length limitation rather than prediction failures.

4.4 Qualitative Analysis

Figure 3 examines semantic interpretation and proposal-conditioned localization on simulated *accident* and native low-light *jenga* (Scheerlinck et al., 2019); additional cases are in Appendix C.

In *accident*, EFG-VLM places the critical interaction near 0.467 s and identifies the white–black car collision, whereas compared systems leave the two-vehicle interaction unresolved. In low-light *jenga*, it identifies a box-shaped object striking the tower at 5.745–6.075 s; GPT-5.5 detects an incoming object without resolving it, while Holmes-VAU and EvQA-FRT mainly describe the sub-

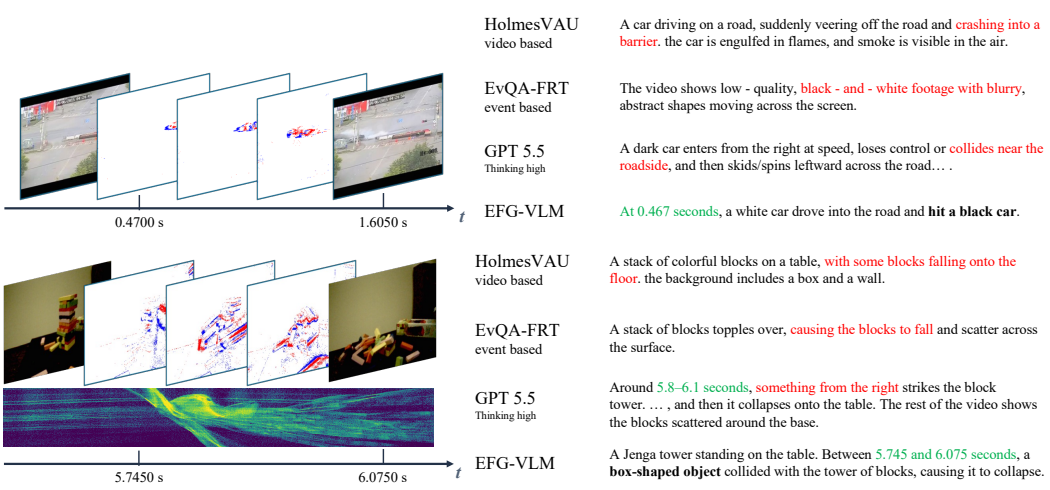


Figure 3: Representative qualitative cases. Top: on the simulated *accident* sequence, EFG-VLM resolves the critical event as a collision between two vehicles. Bottom: on the native low-light *jenga* sequence, it identifies the incoming box-shaped object rather than only the subsequent collapse.

Table 3: Ablation study of EFG-VLM. Detection metrics are reported over the pooled evaluation samples from XD-Violence and UCF-Crime; language metrics evaluate generated anomaly descriptions.

Variant	Ablation setting	Acc.	Macro-F1	BLEU	CIDEr	METEOR	ROUGE-L
Reference and representation baseline							
Full model	None	48.09	16.55	0.064	0.150	0.207	0.239
Event-frame-only baseline	Event frames + Qwen	28.47	12.63	0.033	0.074	0.155	0.182
Component ablations							
w/o spatial graph	Motion graph only	39.96	13.49	0.036	0.079	0.160	0.189
w/o causal motion graph	Spatial graph only	49.80	18.74	0.036	0.085	0.159	0.186
w/o change-aware anchors	Uniform 5-point anchors	46.83	17.02	0.037	0.074	0.163	0.190
w/o coordinate-aware fusion	Global context only	39.43	14.88	0.036	0.080	0.161	0.187
w/o semantic recurrent memory	No local semantic GRU	40.89	13.97	0.035	0.075	0.160	0.188

sequent collapse. Together, the cases suggest that fine-grained event dynamics help recover short causal interactions that remain ambiguous in frame-level observations.

4.5 Ablation Study

Table 3 reveals three complementary trends about how the major components contribute to anomaly understanding.

First, event-native graph evidence provides information that cannot be recovered from sparse event frames alone. Replacing the full representation with the event-frame-only baseline reduces accuracy from 48.09 to 28.47 and CIDEr from 0.150 to 0.074. The simultaneous degradation in recognition and description indicates that semantic event frames alone do not preserve the local temporal evidence required by the complete model.

Second, the spatial and causal motion relations exhibit a clear functional asymmetry. Removing the spatial graph reduces accuracy from 48.09 to 39.96 and CIDEr from 0.150 to 0.079, whereas removing the causal motion graph slightly increases accuracy and Macro-F1 to 49.80 and 18.74 but decreases CIDEr to 0.085 and lowers every other reported language metric. The spatial graph preserves near-simultaneous geometry—event contours, relative layout, and contact structure—that provides stable evidence for recognizing actors and coarse anomaly categories. The causal graph instead emphasizes how those structures evolve from earlier to later events. Such directed dynamics are less necessary for coarse discrimination when structural and visual-semantic cues are already

available, but become important for explaining how an abnormal interaction unfolds. The ablation therefore suggests that recognition relies more strongly on structural evidence, while fine-grained description benefits more from directed motion evolution.

Third, the way local evidence is selected, temporally organized, and aggregated is also important. Replacing the change-aware landmarks with uniformly spaced five-point anchors reduces CIDEr from 0.150 to 0.074, despite only a modest change in recognition performance. This suggests that sampling around internal temporal changes is particularly important for preserving semantically informative moments. Removing coordinate-aware local fusion produces a broader degradation, reducing accuracy to 39.43 and CIDEr to 0.080, indicating that local scene–motion evidence benefits from explicit proposal-relative temporal geometry. Removing the semantic recurrent memory similarly reduces accuracy to 40.89 and Macro-F1 to 13.97, while CIDEr falls to 0.075. Thus, short-range recurrent modeling remains useful after graph propagation for aggregating ordered local dynamics. Overall, the ablations support a hierarchy in which event-native graphs preserve structural/dynamic evidence, change-aware anchors select where it is observed, coordinate-aware fusion grounds it in proposal time, and recurrent aggregation summarizes its local evolution.

5 Conclusion

We present EFG-VLM, combining sparse event-frame semantics with event-native spatial–motion graphs for long-stream anomaly understanding. Proposal-guided shared anchors align semantic and motion evidence, while hierarchical fusion integrates local dynamics with broader temporal context and structured localization priors before language reasoning. Experiments show clear event-source gains in recognition and description, and timestamped simulated/native cases illustrate proposal-conditioned localization of short-lived physical interactions. The current framework prioritizes high-fidelity offline reasoning; future work will investigate transferring this capability to online deployment through causal proposal generation and lightweight distillation while retaining event-native scene–motion evidence.

AI USE STATEMENT

Generative AI tools were used to assist with English-language editing and limited code completion and debugging. All generated code was manually inspected, modified where necessary, and validated through the experiments reported in this paper. The tools were not used to determine the research questions, formulate the proposed method, generate experimental data, or select the reported results. The authors take full responsibility for the manuscript, code, figures, references, and experimental conclusions.

ETHICS STATEMENT

This work studies video anomaly understanding for research purposes using previously released benchmark and event-camera datasets, and does not collect new human-subject data. Video anomaly detection can nevertheless have dual-use implications, particularly in surveillance and other safety-critical settings. Model predictions may contain classification, localization, or language-generation errors, and performance on benchmark datasets does not guarantee reliability across populations, environments, or deployment conditions. The proposed system should therefore not be used as the sole basis for consequential decisions involving individuals. We do not make claims about demographic fairness or deployment readiness, and encourage future work to evaluate robustness, bias, privacy, and human oversight before real-world use.

REPRODUCIBILITY STATEMENT

We provide the model architecture, training procedure, evaluation protocol, and major implementation details in the main paper and appendix. The event benchmark preserves the original HIVAU-70K train/test split and converts each source RGB video one-to-one into an event stream using a fixed v2e configuration. The complete simulation parameters, including the DVS346 sensor setting, timestamp resolution, exposure configuration, contrast thresholds, threshold noise, and cutoff frequency, are specified in the appendix. We also document the construction of MiniDataset, evaluation metrics, baseline protocols, graph construction, structured language interface, and multi-stage optimization. The main training runs use seed 42 with four training processes, and each stage uses the latest saved checkpoint rather than selecting a validation-best checkpoint. Code, configuration files, and preprocessing scripts will be released to facilitate reproduction.

REFERENCES

Zhaolin Cai, Fan Li, Huiyu Duan, Lijun He, and Guangtao Zhai. Steering and rectifying latent representation manifolds in frozen multi-modal LLMs for video anomaly detection. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ol78tbpqyu.

Congqi Cao, Yue Lu, and Yanning Zhang. Context recovery and knowledge retrieval: A novel two-stream framework for video anomaly detection. *IEEE Transactions on Image Processing*, 33: 1810–1825, 2024. doi: 10.1109/TIP.2024.3372466.

Jianwen Cao, Jiaxu Xing, Nico Messikommer, and Davide Scaramuzza. Generative event pretraining with foundation model alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Thomas Dalgaty, Thomas Mesquida, Damien Joubert, Amos Sironi, Pascal Vivet, and Christoph Posch. Hugnet: Hemi-spherical update graph neural network applied to low-latency event-based optical flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 3953–3962, June 2023.

Manon Dampfhoffer, Thomas Mesquida, Damien Joubert, Thomas Dalgaty, Pascal Vivet, and Christoph Posch. Graph neural network combining event stream and periodic aggregation for low-latency event-based vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6909–6918, June 2025.

Yongjian Deng, Hao Chen, Hai Liu, and Youfu Li. A voxel graph cnn for object classification with event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1172–1181, June 2022.

Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2022. doi: 10.1109/TPAMI.2020.3008413.

Or Hirschorn and Shai Avidan. Normalizing flows for human pose anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13545–13554, October 2023.

Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1312–1321, June 2021.

Guoqiang Liang, Kanghao Chen, Hangyu Li, Yunfan Lu, and Lin Wang. Towards robust event-guided low-light image enhancement: A large-scale real-world event-image dataset and novel approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23–33, June 2024.

Shaoyu Liu, Jianing Li, Guanghui Zhao, Yunjian Zhang, and Xiangyang Ji. Eventbench: Towards comprehensive benchmarking of event-based mllms. *arXiv preprint arXiv:2511.18448*, 2025a.

Shaoyu Liu, Jianing Li, Guanghui Zhao, Yunjian Zhang, Xin Meng, Fei Richard Yu, Xiangyang Ji, and Ming Li. Eventgpt: Event stream understanding with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29139–29149, June 2025b.

Shaoyu Liu, Jianing Li, guanghui zhao, Yunjian Zhang, Wen Jiang, Ming Li, and Xiangyang Ji. Eventflash: Towards efficient MLLMs for event-based vision. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=QuvGqzLwf6.

Zihao Liu, Xiaoyu Wu, Jianqin Wu, Xuxu Wang, and Linlin Yang. Language-guided open-world video anomaly detection under weak supervision. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=70lFIQG5OB.

Hanyue Lou, Jiayi Zhou, Yang Zhang, Boyu Li, Yi Wang, Ye Guangnan, and Boxin Shi. Reconstruction as a bridge for event-based visual question answering. In *arXiv preprint arXiv:2512.11510*, 2025.

Yuanbin Qian, Shuhan Ye, Chong Wang, Xiaojie Cai, Jiangbo Qian, and Jiafei Wu. Ucf-crime-dvs: A novel event-based dataset for video anomaly detection with spiking neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(6):6577–6585, Apr. 2025. doi: 10.1609/aaai.v39i6.32705. URL https://ojs.aaai.org/index.php/AAAI/article/view/32705.

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.

Simon Schaefer, Daniel Gehrig, and Davide Scaramuzza. Aegnn: Asynchronous event-based graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12371–12381, June 2022.

Cedric Scheerlinck, Henri Rebecq, Timo Stoffregen, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Ced: Color event camera dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Jiaqi Tang, Hao Lu, Ruizheng Wu, Xiaogang Xu, Ke Ma, Cheng Fang, Bin Guo, Jiangbo Lu, Qifeng Chen, and Ying-Cong Chen. Hawk: Learning to understand open-world video anomalies. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 139751–139785. Curran Associates, Inc., 2024. doi: 10.52202/079017-4435. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/fca83589e85cb061631b7ebc5db5d6bd-Paper-Conference.pdf.

Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16155–16164, June 2021.

Wenhao Xu, Xin Dong, Yue Li, Haoyuan Shi, and Zhiwei Xiong. Eventstu: Event-guided efficient spatio-temporal understanding for video large language models. *arXiv preprint arXiv:2511.18920*, 2025.

Shengtian Yang, Yue Feng, Yingshi Liu, Jingrou Zhang, and Jie Qin. Monitor: Exploiting large language models with instruction for online video anomaly detection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=6Had86RHix.

Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, and Nong Sang. Holmes-vau: Towards long-term video anomaly understanding at any granularity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13843–13853, June 2025.

Hanyu Zhou and Gim Hee Lee. Llafea: Frame-event complementary fusion for fine-grained spatiotemporal understanding in lmms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22294–22304, October 2025.

A Additional Method Details

A.1 Proposal Generation and Anchor Selection

The 4-Hz timeline is constructed from non-overlapping 0.25-second bins, each represented by a polarity-aware spatial histogram and lightweight event statistics. A temporal convolutional encoder with symmetric dilated convolutions processes overlapping blocks, whose outputs are folded into a complete sequence $\{\mathbf{h}_m\}_{m=1}^{M}$. Because the convolutions are symmetric, each timeline feature incorporates both past and future context within a bounded temporal receptive field. With kernel size 3 and dilations $(1, 2, 4, 8, 16, 32)$, the theoretical receptive field is 127 timeline bins, corresponding to 31.75 s at 4 Hz.

The locator predicts eventness, temporal boundaries, centers, and offsets. Candidate intervals are combined with event-rate proposals, boundary-refined, ranked by predicted IoU quality, and filtered by temporal non-maximum suppression. Although the temporal feature context is bounded, proposal generation and ranking are performed over the folded timeline and are therefore not strictly causal. The current framework consequently targets offline long-stream anomaly understanding rather than real-time early-warning detection.

For the two change-aware landmarks, the proposal is divided at its center and the strongest feature change is selected independently from each half. This half-wise selection discourages both adaptive landmarks from concentrating on the same local transition and preserves coverage of the early and late proposal phases. If a valid distinct maximum is unavailable in either half, the corresponding quarter or three-quarter position is used as a fallback. The resulting five-anchor layout therefore combines three deterministic structural references with two content-adaptive internal observations while keeping the number of event-frame and graph tokens fixed across proposals.

A.2 Graph Construction and Local Temporal Aggregation

For each bounded event window, raw events are partitioned into fixed-duration chunks and retained individually up to a fixed node budget. If a chunk exceeds the budget, the temporally ordered event sequence is uniformly subsampled rather than compressed into supernodes. Each retained event is represented by

$$\mathbf{x}_i = [\hat{x}_i, \hat{y}_i, \bar{t}_i, p_i], \tag{14}$$

where $(\hat{x}_i, \hat{y}_i)$ are normalized image coordinates, $\bar{t}_i \in [0, 1]$ is the chunk-local normalized timestamp, and p_i denotes event polarity. Unless otherwise specified, all experiments use the graph-construction settings in Table 4.

Table 4: Default graph-construction hyperparameters used in our experiments. Spatial radii are measured in normalized image coordinates.

Parameter	Symbol	Value
Graph chunk rate	–	120 Hz
Maximum nodes per chunk	$N_{\max}$	12,000
Spatial micro-temporal bins	B_s	2
Spatial KNN neighbors	K_s	16
Spatial radius	R_s	0.04
Motion history chunks	H_m	5
Motion temporal window	τ_m	0.04 s
Motion KNN neighbors	K_m	12
Motion radius	R_m	0.04
Motion candidate multiplier	–	4

Spatial graph. The spatial relation is intended to capture local structure among near-simultaneous events. Each chunk is divided into B_s micro-temporal bins according to

$$b_i = \min\left(B_s - 1, \lfloor B_s \bar{t}_i \rfloor\right). \tag{15}$$

Let κ_i denote the chunk index of event i and $\mathbf{r}_i = (\hat{x}_i, \hat{y}_i)$ its normalized image-plane position. Spatial neighbors are searched only among events belonging to the same local stream, chunk, and micro-temporal bin:

$$\mathcal{E}^{\text{spa}} = \{(i, j) \mid \kappa_i = \kappa_j,\ b_i = b_j,\ i \in \text{KNN}_{xy}(j),\ d_{ij} \le R_s\}, \tag{16}$$

where

$$d_{ij} = \|\mathbf{r}_j - \mathbf{r}_i\|_2 \tag{17}$$

is the normalized image-plane distance and R_s is the spatial radius. Self-edges are removed after the nearest-neighbor search. Restricting the spatial relation to near-simultaneous events prevents temporally separated observations from being treated as the same local structure.

Causal motion graph. The motion relation instead connects temporally ordered events and is allowed to cross chunk boundaries. For each target event j, candidate predecessors are collected from its current chunk and a bounded number of preceding chunks within the same local stream. An edge is retained only if

$$(i, j) \in \mathcal{E}^{\text{mot}} \quad \Longrightarrow \quad 0 < t_j - t_i \le \tau_m, \qquad d_{ij} \le R_m, \tag{18}$$

where t_i and t_j are absolute event timestamps, τ_m is the maximum temporal gap, and R_m is the motion search radius. Image-plane KNN first provides a bounded candidate set, after which candidates are filtered by the temporal and spatial constraints above. The remaining predecessors for each target are ranked by

$$q_{ij} = d_{ij} + \lambda_p |p_j - p_i|, \qquad \lambda_p = 0.02, \tag{19}$$

and the lowest-scoring neighbors are retained. This ranking favors spatially close predecessors while applying a small penalty to polarity disagreement. Unlike the spatial relation, all motion edges therefore propagate information strictly from earlier to later events.

Edge descriptor. Both relations use the same eight-dimensional relative edge descriptor

$$\mathbf{a}_{ij} = [\Delta x_{ij}, \Delta y_{ij}, \Delta t_{ij}, d_{ij}, v_{ij}^x, v_{ij}^y, \gamma_{ij}^p, \gamma_{ij}^c]. \tag{20}$$

For an edge from source event i to target event j, the relative displacement and elapsed time are defined as

$$\Delta x_{ij} = \hat{x}_j - \hat{x}_i, \qquad \Delta y_{ij} = \hat{y}_j - \hat{y}_i, \qquad \Delta t_{ij} = t_j - t_i, \tag{21}$$

with

$$d_{ij} = \sqrt{(\Delta x_{ij})^2 + (\Delta y_{ij})^2}. \tag{22}$$

The local image-plane velocity proxy is computed from the absolute temporal separation:

$$v_{ij}^x = \mathrm{clip}\left(\frac{\Delta x_{ij}}{\max(|\Delta t_{ij}|, \epsilon_t)}, -16, 16\right), \qquad v_{ij}^y = \mathrm{clip}\left(\frac{\Delta y_{ij}}{\max(|\Delta t_{ij}|, \epsilon_t)}, -16, 16\right), \tag{23}$$

where $\epsilon_t = 10^{-4}$ prevents numerical instability for nearly simultaneous events. The absolute value in the denominator allows the same descriptor definition to be used for both spatial and causal motion relations, while the signed Δt_{ij} remains explicitly available as a separate descriptor dimension.

Polarity agreement is encoded by

$$\gamma_{ij}^p = 1 - \frac{\min(|p_i - p_j|, 2)}{2}, \tag{24}$$

so that larger values indicate more consistent event polarities. We further define a relation confidence that jointly reflects spatial proximity and polarity consistency:

$$\gamma_{ij}^c = \exp\left(-\frac{d_{ij}}{\max(R_r, \epsilon_r)}\right)\left(\frac{1 + \gamma_{ij}^p}{2}\right), \tag{25}$$

where $R_r = R_s$ for spatial edges and $R_r = R_m$ for motion edges, and $\epsilon_r = 10^{-6}$ prevents division by zero. Thus, γ_{ij}^c assigns larger values to spatially closer edges with stronger polarity agreement.

Relation-specific message passing. Because the spatial and motion graphs represent different interactions, they are processed independently. For relation $r \in \{s, m\}$, source-to-target messages are conditioned on the corresponding edge descriptor:

$$\mathbf{m}_{ij}^r = \mathbf{W}_r \mathbf{h}_i + \psi_r(\mathbf{a}_{ij}), \tag{26}$$

where ψ_r projects the eight-dimensional edge descriptor into the node-feature space. Incoming messages are mean-aggregated at each target node and followed by a residual transformation. The spatial and motion representations are subsequently combined by the node-wise gate described in the main paper. For nodes without an incoming motion edge, the motion contribution is masked out.

Local temporal aggregation. Graph propagation preserves event-level local relations, but each bounded window may still contain multiple temporally ordered chunks. We first aggregate node features within each chunk using complementary mean, max, and gated-attention pooling. The resulting chunk representations form an ordered local sequence that is processed by a causal multi-scale temporal encoder to capture evolution over multiple temporal ranges. A short GRU then models semantic progression within the bounded window. Its hidden state is initialized to zero for every window and is never propagated across windows, proposals, or videos. Finally, salient pooling over the local chunk sequence produces the landmark-level motion representation $\mathbf{m}_{p,l}$ used by the hierarchical scene–motion encoder.

A.3 Context and Language Interface

Within each proposal, the five projected visual landmarks are pooled to form proposal-level visual context and then pooled across proposals to form a global visual representation $\mathbf{C}_V$. The folded 4-Hz timeline is pooled by mean, max, and learned attention to obtain $\mathbf{C}_T$. The global token is formed by residual visual injection,

$$\mathbf{C} = \mathbf{C}_T + g_v \mathbf{C}_V, \qquad g_v = \sigma(\phi_v[\mathbf{C}_T; \mathbf{C}_V]). \tag{27}$$

The numerical temporal prior explicitly carries localization evidence from the preceding stages into the language model. Its first 48 dimensions are three 16-bin traces from the Stage-1 locator—eventness, start, and end—and the remaining six scalars are the proposal score, validity, description readiness, anomaly probability, and refined start/end coordinates. This yields the 54-dimensional prior $\boldsymbol{\pi}_p \in \mathbb{R}^{54}$. The Stage-1 proposal therefore serves both as the coordinate frame for the final start/end bins and as an explicit source of temporal evidence available to Qwen through the prior token. Video context, proposal summaries, landmark features, temporal coordinates, and priors are projected independently into the Qwen hidden space. Token-type embeddings distinguish video, summary, the five landmark roles, and prior tokens, while proposal-slot embeddings preserve event identity.

A.4 Multi-Stage Optimization

Stage 1 trains the temporal locator using class-agnostic activity boundaries. In Stage 2, the locator is fixed and the scene–motion context encoder is optimized using proposal-validity, description-readiness, boundary, anomaly-prior, and semantic objectives. Semantic supervision is obtained from frozen Qwen text embeddings of the available video, event, and clip descriptions. In Stage 3, the localization and context encoders are frozen; contextual video representations, proposal summaries, landmark tokens, coordinates, and priors are cached before training the multimodal bridge and Qwen LoRA adapters.

B Additional Experimental Details

B.1 Data Construction and Evaluation Protocol

HIVAU-70K inherits its train/test partition, annotations, and instruction data from Holmes-VAU. Each source RGB video is converted one-to-one into an event stream using v2e (Hu et al., 2021), while preserving the original split and annotations. All videos are simulated with the same fixed configuration, and no frame interpolation is applied during conversion. The complete simulation parameters are listed in Table 5.

Table 5: v2e simulation parameters used for all HIVAU-70K source videos.

Parameter	Value
Sensor preset	DVS346
Timestamp resolution	0.003 s
Automatic timestamp resolution	Disabled
Frame interpolation	Disabled
DVS exposure mode	Duration
DVS exposure duration	0.005 s
Positive contrast threshold	0.15
Negative contrast threshold	0.15
Threshold noise std.	0.03
Cutoff frequency	10 Hz

After conversion, the simulated event stream remains in one-to-one correspondence with its source RGB video and therefore reuses the original HIVAU-70K annotations and instruction data. After merging the XD-Violence and UCF-Crime source subsets, the training and test splits contain 160,509 and 33,518 instructions, respectively.

MiniDataset is constructed only from the HIVAU-70K evaluation split and introduces no separate training data. Its samples are grouped according to the actual clip duration rather than the annotated anomaly duration. The five duration ranges are $<$ 1s, 1–5s, 5–10s, 10–20s, and $>$ 20s. The $<$ 1s group contains 78 samples, while each of the remaining four groups contains 372 samples, yielding 1,566 samples in total. The duration bins contain different clips rather than repeated versions of the same videos, so trends are interpreted as robustness across regimes rather than causal effects of clip length. EvQA-ART remains below EvQA-FRT across all duration bins. A plausible method-specific factor is that ART uses activity-triggered sparse reconstruction, whose region-level triggering depends on local event-density statistics; these statistics may shift under v2e simulation,

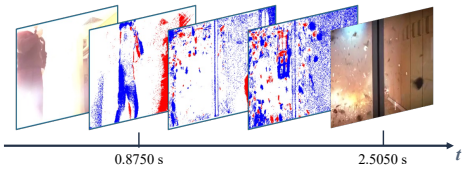
whereas FRT reconstructs dense frames at a fixed rate. We therefore treat this simulation-to-trigger interaction as a hypothesis rather than a measured failure mechanism. EventGPT's missing long-duration entries reflect its input-length limitation rather than prediction failures.

For GPT-5.5, we retain the same anomaly prompt used in the dataset evaluation and tile the sampled I/P/B RGB frames into a single composite image for one API request. MoniTor (Yang et al., 2026) is evaluated using the API pipeline provided by its official repository. API refusals and outputs that cannot be parsed into the required prediction format are counted as incorrect predictions. Because the training corpora of online closed-source APIs are undisclosed, possible prior exposure to public benchmark videos cannot be excluded.

The proposal locator operates on a 4-Hz temporal timeline, while the graph branch processes the raw event stream using 120-Hz chunks. The language backbone is Qwen 3.5-2B with LoRA adaptation. Main-chain checkpoints are trained with seed 42 using four training processes. For each stage, we use the latest saved checkpoint rather than selecting the best checkpoint according to a validation metric.

C Additional Qualitative Results

The simulated *explosion* and *fight* cases are selected from the HIVAU-70K evaluation split and converted with the same v2e pipeline as the quantitative benchmark. The native *gnome* sequence comes from *High Speed and High Dynamic Range Video with an Event Camera* (Rebecq et al., 2021), and *waterbomb* comes from Time Lens (Tulyakov et al., 2021). Because the simulated cases are not native synchronized RGB–event recordings, continuous ROI projections are shown only for the native examples.



HolmesVAU
video based

A series of images depicting a building engulfed in flames and smoke, with debris flying through the air. the scene is chaotic, … .

EvQA-FRT
event based

… a window breaks, revealing a building engulfed in flames and smoke, with debris and collapsing structures.

GPT 5.5
Thinking high

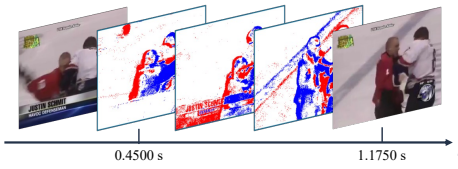
…. A sudden blast fills the corridor with bright light, smoke, sparks, and debris. … . Later shots show debris and sparks … .

EFG-VLM

At around 0.878 seconds, A **large explosion occurring** in an urban area, with thick smoke and fire billowing into the sky. the explosion is… .

Figure 4: Additional v2e-simulated HIVAU-70K example: *explosion*.

In *explosion*, several models recognize fire, smoke, or a blast, while EFG-VLM additionally anchors the event near 0.878 s.



HolmesVAU
video based

Two ice hockey players engaged in a physical altercation on the ice rink. … . they are grappling with each other, pushing and shoving, and attempting to gain control of the situation.

EvQA-FRT
event based

… a person in a dark suit with a nameplate reading "JUSTIN SCHMIDT" and "BANK OF CHINA" visible, suggesting a professional setting (likely a bank or office).

GPT 5.5
Thinking high

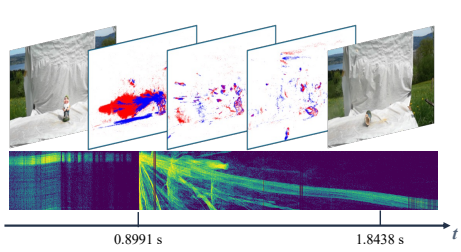
One player in a red jersey and one in a white jersey grab each other's jerseys while skating/standing on the ice. … the player in red **throws several punches** …

EFG-VLM

A hockey game where a player in a white jersey with the number 8 is **being tackled by an opponent** in a black jersey. Around the 0.453 seconds, … the player in black is attempting to lift him up.

Figure 5: Additional v2e-simulated HIVAU-70K example: *fight*.

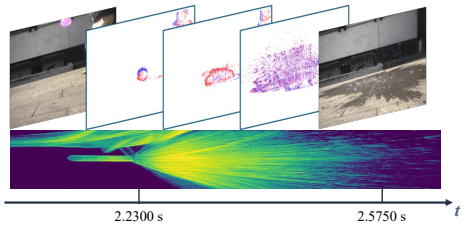
In *fight*, EFG-VLM localizes the relevant interaction near 0.453 s and describes the tackle/lifting motion, whereas EvQA-FRT produces a scene-level hallucination.



HolmesVAU video based	A … person, sitting on a white sheet of paper. the figure is wearing a green hat and appears to be engaged in an activity, possibly reading or writing. the background is ….
EvQA-FRT event based	A person stands in front of a large, wrinkled fabric backdrop. The fabric begins to move and ripple, with the person interacting with it as … .
GPT 5.5 Thinking high	At about 1.1 seconds, it is suddenly hit by something from the left side or by a fast impact. A dust/smoke cloud and fragments appear, and pieces fly … .
EFG-VLM	A goblin was placed on the table; between 1.152 seconds and 2.069 seconds, an **explosion occurred** above his head.

Figure 6: Native open-world RGB–event example: *gnome*.

In *gnome*, EFG-VLM localizes the sudden change and describes an explosion above the figure's head, while GPT-5.5 detects a fast impact and resulting dust/smoke but remains less specific.



HolmesVAU video based	A person standing on a paved surface, wearing … . the person then kicks a small object, causing it to fly into the air and scatter across the ground.
EvQA-FRT event based	… Wait, maybe the video is a sequence of events: (so the person is moving), then a splash (maybe water), …. . Let' a pen (or object) on a gray background, then a person's legs s check the frames again.….. (failure to generate EOS token)
GPT 5.5 Thinking high	The balloon is released or thrown downward near the person's foot. It hits the ground and bursts, producing a large splash of water … .
EFG-VLM	A man dropped a water balloon; around 2.23 seconds, the water balloon **hit the ground and burst**.

Figure 7: Native open-world RGB–event example: *waterbomb*.

In *waterbomb*, both GPT-5.5 and EFG-VLM recognize the balloon burst, while EFG-VLM explicitly localizes the impact near 2.23 s. Holmes-VAU instead interprets the action as kicking a small object, and EvQA-FRT fails to terminate normally.