TFE-GRAPHRAG: TEMPORAL-FREQUENCY EVIDENCE RETRIEVAL FOR EFFICIENT EVENT-BASED MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Event-based multimodal large language models (MLLMs) struggle to process event streams with microsecond-level temporal resolution. Existing event-based algorithms typically divide event streams into fixed groups, potentially discarding crucial temporal information across different motion speeds. While retrieval-augmented generation (RAG) offers a promising remedy by organizing query-relevant visual evidence into a compact context, most existing methods flatten videos into independent segments, breaking their inherent spatiotemporal structure. In this work, we propose TFE-GraphRAG, a temporal-frequency event GraphRAG framework to enhance event stream understanding. Our approach introduces two key innovations: (i) it uses continuous wavelet transform (CWT) multiresolution analysis to adaptively slice event streams, preserving temporal information; and (ii) it introduces an event evidence graph that preserves semantic relationships across time, space, and frequency, and uses frequency prior knowledge to improve retrieval across different motion scenarios. We evaluate our framework with multiple open-source MLLMs on two event-stream understanding benchmarks. Our approach improves overall performance by 2.3–4.8 points over the state-of-the-art event MLLM methods. Our code is available at the Supplementary Materials.

1 INTRODUCTION

Event cameras asynchronously emit timestamped polarity changes instead of sampling complete images at a fixed frame rate (Posch et al., 2014; Gallego et al., 2022). Their microsecond-scale resolution, high dynamic range, low latency, and sparse output support recognition, reconstruction, and segmentation during rapid motion and difficult illumination (Lagorce et al., 2017; Rebecq et al., 2019; Alonso & Murillo, 2019). Event-based vision systems typically convert raw event streams into structured representations for downstream processing.

Multimodal large language models (MLLMs) support scene understanding, visual question answering, and instruction following over frame sequences (Li et al., 2023; Liu et al., 2023; Huang et al., 2024; Qian et al., 2024). However, their image-oriented interfaces cannot directly process asynchronous event streams. For example, EventGPT (Liu et al., 2025) converts event streams into dense token sequences for language modeling, whereas EventVL (Li et al., 2025) integrates RGB frames with event data to enhance multimodal reasoning. LLaFEA (Zhou & Lee, 2025) fuses frame and event features for fine-grained spatiotemporal understanding, while EventFlash (Liu et al., 2026) combines adaptive temporal merging with density-guided spatial sparsification for efficient event processing.

Common event-stream partitioning methods use either fixed time intervals or fixed event counts. These predetermined strategies cannot adapt to scene dynamics and process the entire frame uniformly, thereby ignoring spatial locality. Retrieval-augmented generation (RAG) and adaptive keyframe sampling offer promising approaches to long-video processing by selecting a small subset of segments or frames that fits within a limited context window (Ren et al., 2025). However, flattening a video into independent segments disrupts its intrinsic spatiotemporal structure and separates

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.

contextual evidence that should remain connected (Tang et al., 2025). Together, these limitations expose two coupled problems: temporal information loss and spatiotemporal context fragmentation.

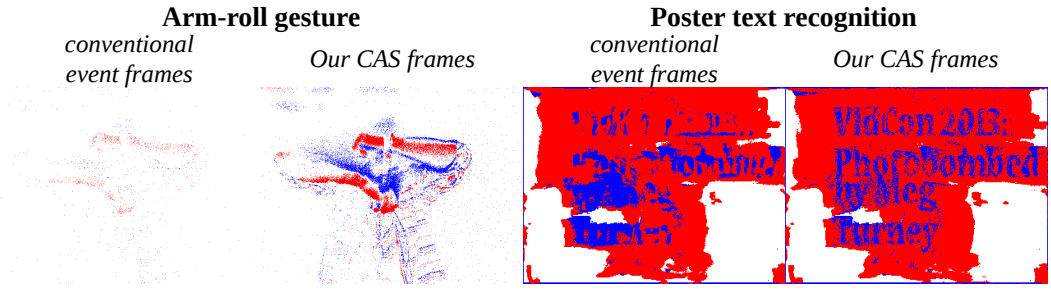

Figure 1: Comparison between conventional event frames and our CWT Adaptive Slicing (CAS) frames; red and blue denote opposite event polarities. For the arm-roll gesture, our low-frequency CAS frame aggregates the sustained motion over a longer adaptive interval, revealing clearer arm-and-torso structure. For poster-text recognition under a changing background, our high-frequency CAS frame preserves rapid character-edge responses and keeps the text on the poster more legible.

Temporal information loss. To address this problem, we introduce CWT Adaptive Slicing (CAS) frames. CAS applies CWT directly to event streams and uses patch- and frequency-specific energy flux to determine adaptive temporal intervals. For each interval index and frequency group, the corresponding CWT coefficient blocks from all spatial patches are assembled into a frequency-specific CAS frame. This preserves patch-local temporal adaptation while providing a frame-level evidence unit for graph construction. Figure 1 illustrates how low-frequency CAS frames consolidate sustained motion, whereas high-frequency CAS frames retain rapid edge changes.

Spatiotemporal context fragmentation. To address the limitations of segment-based RAG, we construct a structured frame-level event evidence graph. Each node represents one frequency-specific CAS frame and stores its metadata and normalized visual embedding, while the frame tensor remains in external storage. Temporal edges connect consecutive frames within the same frequency group, frequency edges connect corresponding intervals across groups, and spatial edges link frames with similar visual embeddings. Graph construction is performed offline and is query independent, allowing the same graph to support multiple questions without reprocessing the event stream. Query-Conditioned Evidence Retrieval (QCER) identifies query-relevant roots and performs bounded graph expansion over the three edge types. It returns compact relational evidence and loads only the selected CAS frames for MLLM reasoning, with the retrieval budget adapted to the input length.

We evaluate the resulting interface on two event-stream understanding benchmarks using Qwen2.5 MLLMs at multiple scales. Our experiments compare task quality and inference efficiency against dense event processing and existing event-stream methods. On EventMind, TFE-GraphRAG outperforms EventFlash on all four metrics at both model scales, with gains of 2.3–4.8 points.

In summary, our contributions are:

- CAS frames, which apply CWT directly to polarity-separated event streams, adapt patch-level temporal boundaries using frequency-specific energy flux, and assemble the resulting CWT coefficient blocks into frequency-specific frames;
- an offline, reusable frame-level event evidence graph whose temporal, frequency, and spatial edges preserve chronological, cross-frequency, and spatial-layout relationships;
- Query-Conditioned Evidence Retrieval (QCER), which combines semantic and frequency compatibility with bounded multi-hop graph expansion and materializes only the selected CAS frames for MLLM reasoning.

2 Related Work

2.1 Event-stream MLLMs.

EventGPT (Liu et al., 2025) couples event features to a language model and introduces EventChat for instruction-following evaluation, while EventVL (Li et al., 2025) develops a generative MLLM for event-stream understanding. LLaFEA (Zhou & Lee, 2025) fuses complementary frame and event cues. Long-stream systems increasingly compress evidence before language decoding. EventFlash (Liu et al., 2026) adaptively aggregates temporal windows and uses density-guided sparse attention. These studies establish event-language alignment and selective compression, but their model inputs are reconstructed frames, dense features, or compact token sequences formed before generation. Accordingly, they do not explicitly separate persistent evidence construction from question-conditioned relational composition.

2.2 Event evidence representation.

Methods that adapt event aggregation to scene dynamics directly address temporal information loss. TORE volumes (Baldwin et al., 2023) retain recent per-pixel event times without fixed time blocking. SpikeSlicer (Cao et al., 2024) instead uses a spiking neural network as an adaptive trigger and refines its slicing decisions using feedback from a downstream model. Adaptive Spatial-Temporal Window (ASTW) (Sui et al., 2026) derives patch-level temporal windows from event density, combining temporal adaptivity with spatial locality for heterogeneous-velocity scenes. Wavelet representations provide a complementary multiscale treatment: Cheng et al. (2025) combine temporal and spatial wavelet transforms for recognition. These approaches preserve local timing or adapt aggregation boundaries, but their representations primarily support immediate downstream inference rather than persistent, query-addressable evidence for later question-conditioned reasoning.

2.3 Graph retrieval for relational evidence composition.

Video-specific RAG increasingly uses graph structure to preserve relationships beyond independently retrieved clips. VideoRAG (Ren et al., 2025) combines graph-based textual knowledge grounding with multimodal retrieval for extremely long videos. Vgent (Shen et al., 2025) represents clips with semantic graphs and introduces structured verification to filter retrieval noise and aggregate evidence. VideoStir (Fu et al., 2026) constructs a spatiotemporal clip graph and performs multi-hop, intent-aware retrieval across distant events. G2F-RAG (Yang et al., 2026) builds an offline video knowledge graph, retrieves a minimal subgraph, and renders it as a single reasoning frame for visual-space fusion. These approaches establish structured video retrieval, but their graphs rely on RGB clips, text-derived entities, or agent-generated knowledge. An event evidence graph must instead be constructed from uncertain temporal and motion cues, where an aggressive edge can turn proximity into an unsupported identity or causal claim. Its evidence units and relations must therefore be inferred conservatively from temporal, spatial, and relative-frequency cues.

3 Method

TFE-GraphRAG is a temporal-frequency GraphRAG framework for event-based MLLMs that preserves critical event dynamics while reducing unnecessary visual tokens. Figure 2 shows its two modules: offline Wavelet Event Memory Construction (WEMC) and online Query-Conditioned Evidence Retrieval (QCER). WEMC applies CWT to polarity-separated raw events, uses flux-guided CAS to construct frequency-specific frames, and represents each frame as one graph node containing metadata and a normalized visual embedding. Temporal, frequency, and spatial edges organize the frame-level evidence. For each query, QCER estimates a soft frequency preference, selects three roots using semantic and frequency compatibility, and expands them over sufficiently strong edges. Only the retrieved CAS frames are loaded and encoded for the MLLM.

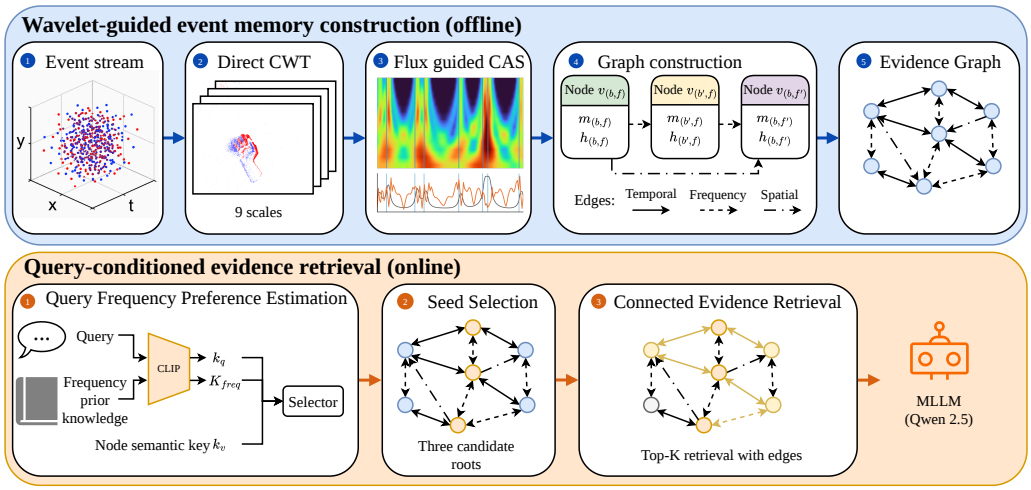

Figure 2: Overview of TFE-GraphRAG. Offline WEMC converts an event stream into frequency-specific CAS frames and constructs a reusable frame-level graph with temporal, frequency, and spatial edges. Online QCER uses semantic and frequency compatibility to select three roots, then performs weight-guided multi-hop expansion and loads the retrieved CAS frames for MLLM reasoning.

3.1 Wavelet Event Memory Construction

3.1.1 CWT Adaptive Slicing Frame Generation

An event stream is $\mathcal{E} = \{e_i\}_{i=1}^{N}$, where $e_i = (x_i, y_i, t_i, p_i)$ gives pixel location, timestamp, and polarity $p_i \in \{-1, +1\}$. For each polarity, these raw events define an impulse field

$$E_p(x, y, t) = \sum_{i:p_i=p} \delta(x - x_i)\delta(y - y_i)\delta(t - t_i). \tag{1}$$

We evaluate a Mexican-hat CWT at output centers τ:

$$W_{p,s}(\tau, x, y) = \frac{1}{\sqrt{s}} \int E_p(x, y, t)\psi\left(\frac{t - \tau}{s}\right) dt, \tag{2}$$

where $\psi(u) = \frac{2}{\sqrt{3}\pi^{1/4}}(1 - u^2)e^{-u^2/2}$.

Output centers are sampling points rather than input bins. We use nine logarithmically spaced Mexican-hat scales at three steps per octave. All nine scales are evaluated on a common grid of 256 output centers, with the two polarities retained separately.

The scale-dependent resolution of the CWT is useful for nonstationary event streams. Small scales provide short support for brief transients, whereas large scales provide longer support for sustained activity. This multi-support representation lets a variable-duration bin be interpreted against the sampled support closest to its length. Because smaller CWT scales indicate higher relative frequencies, we order the nine scales from smallest to largest and partition them into $\mathcal{S}^H = \{1, 2, 3\}$, $\mathcal{S}^M = \{4, 5, 6\}$, and $\mathcal{S}^L = \{7, 8, 9\}$. Let $\mathcal{F} = \{H, M, L\}$; throughout the graph, $f \in \mathcal{F}$ is one discrete relative-frequency label.

CAS is applied separately to each 4×4-pixel spatial patch and frequency group. Different patch–frequency pairs may therefore use different temporal boundaries. For patch r and frequency group f, the CWT energy at output center τ is

$$e_{r,f,\tau} = \sum_{p\in\{-1,+1\}} \sum_{s\in\mathcal{S}^f} \sum_{(x,y)\in r} |W_{p,s}(\tau, x, y)|^2, \qquad f \in \mathcal{F}. \tag{3}$$

Here $\tau \in \{1, \ldots, 256\}$ indexes consecutive CWT output centers, and $\tau - 1$ denotes the immediately preceding center. The frequency-specific energy flux is

$$F_{r,f,\tau} = |e_{r,f,\tau} - e_{r,f,\tau-1}| . \tag{4}$$

High flux indicates a strong energy change within frequency group f. For each patch-frequency pair (r, f), CAS first selects $B - 1$ high-flux positions as initial boundaries and then refines each boundary toward a nearby peak of $e_{r,f,\tau}$. Thus the fixed-bin split and CAS use exactly the same number of intervals; only their boundary locations change. Although the resulting CWT coefficient blocks may have different start and end times across patches, every patch-frequency pair contains the same ordered number B of blocks. Consequently, for each frequency group f and interval index b, the bth CWT coefficient blocks from all spatial patches can be assembled into one CAS frame, yielding B CAS frames per frequency group. For one event stream, WEMC produces $3B$ CAS frames, with B interval indices for each of three frequency groups.

3.1.2 Event evidence graph construction

Let $X_{b,f}$ denote the complete CAS frame at interval index b and frequency group f. WEMC creates one frame-level node for each $X_{b,f}$:

$$v_{b,f} = (m_{b,f}, h_{b,f}), \qquad b = 1, \ldots, B, \quad f \in \mathcal{F}. \tag{5}$$

Each node contains exactly two fields: CAS-frame metadata $m_{b,f}$ and a frame embedding $h_{b,f}$. The metadata contains the storage path $\pi_{b,f}$, interval index b, frequency label f, and patchwise temporal-support map $\{(t^0_{r,b,f}, t^1_{r,b,f})\}_{r=1}^{R}$. During offline graph construction, WEMC loads $X_{b,f}$, encodes it with the CLIP visual encoder E_{vis}, and stores the normalized pooled embedding

$$h_{b,f} = \frac{E_{\text{vis}}(X_{b,f})}{\|E_{\text{vis}}(X_{b,f})\|_2}. \tag{6}$$

The graph contains temporal, frequency, and spatial edges over frame-level nodes. Directed temporal edges connect consecutive CAS frames within the same frequency group,

$$\mathcal{R}_{\text{T}} = \{(v_{b,f}, v_{b+1,f}) \mid b = 1, \ldots, B-1,\ f \in \mathcal{F}\}. \tag{7}$$

Frequency edges connect frames with the same interval index across different frequency groups,

$$\mathcal{R}_{\text{F}} = \{(v_{b,f}, v_{b,f'}) \mid b = 1, \ldots, B,\ f \neq f'\}. \tag{8}$$

Spatial edges connect frame pairs according to their frame-level embedding similarity,

$$\mathcal{R}_{\text{Sp}} = \{(v_i, v_j) \mid w^{\text{Sp}}_{ij} = \max(0, \cos(h_i, h_j))\}, \tag{9}$$

where h_i and h_j are the normalized CLIP embeddings of CAS frames i and j, and the maximum operation removes negative similarities. Together, $\mathcal{R} = \mathcal{R}_{\text{T}} \cup \mathcal{R}_{\text{F}} \cup \mathcal{R}_{\text{Sp}}$ and the frame-level nodes $\mathcal{V}$ form the event evidence graph $\mathcal{G} = (\mathcal{V}, \mathcal{R})$.

3.2 Query-Conditioned Evidence Retrieval

3.2.1 Query Frequency Preference and Seed Selection

The CLIP text encoder maps query q to a normalized query key k_q. The same encoder maps three fixed frequency prompts to normalized embeddings, stacked as the rows of K_{freq}. These prompts encode soft rather than deterministic associations: queries about abrupt onsets, offsets, impacts, or direction reversals favor high-frequency evidence; queries about sustained gestures, locomotion, or multi-step action progression favor mid-frequency evidence; and queries about object presence, spatial layout, or slowly varying scene context favor low-frequency evidence, with mid-frequency weight retained when recognition depends on motion-defined contours. Their similarity produces the query's soft frequency preference

$$k_q = \frac{E_{\text{text}}(q)}{\|E_{\text{text}}(q)\|_2}, \qquad \boldsymbol{\alpha}_q = \text{softmax}(K_{\text{freq}} k_q), \tag{10}$$

where $\boldsymbol{\alpha}_q = [\alpha^H_q, \alpha^M_q, \alpha^L_q]$ gives the different frequency preferences. For frame-level node $v_{b,f} = (m_{b,f}, h_{b,f})$, the frequency label f is read from the CAS-frame metadata $m_{b,f}$, and the

stored normalized embedding $h_{b,f}$ is used directly for semantic matching. Semantic and frequency compatibility are

$$\begin{aligned} s_{\text{sem}}(v_{b,f}, q) &= h_{b,f}^{\top} k_q, \\ s_{\text{freq}}(v_{b,f}, q) &= \alpha_q^f, \\ U_{\text{seed}}(v_{b,f} \mid q) &= 0.80 s_{\text{sem}}(v_{b,f}, q) + 0.20 s_{\text{freq}}(v_{b,f}, q). \end{aligned} \tag{11}$$

The three frame-level nodes with the largest U_{seed} values become candidate roots. This is a global top-three operation: it does not force one root from every frequency group or automatically include the frequency siblings of a selected root.

3.2.2 Connected Evidence Retrieval and Interface

Let $\mathcal{V}_{\text{root}}$ contain the three candidate roots. To recover broader context, QCER expands these roots over the temporal, frequency, and spatial links in $\mathcal{G}$. For an edge from u to $v = v_{b,f}$, its query-conditioned weight is

$$W_q(u,v) = \lambda_{\text{T}} \mathbf{1}[(u,v) \in \mathcal{R}_{\text{T}}] + \lambda_{\text{F}} \alpha_q^f \mathbf{1}[(u,v) \in \mathcal{R}_{\text{F}}] + \lambda_{\text{Sp}} w_{uv}^{\text{Sp}} \mathbf{1}[(u,v) \in \mathcal{R}_{\text{Sp}}], \tag{12}$$

where $\mathbf{1}[\cdot]$ is the indicator function and $\lambda_{\text{T}}, \lambda_{\text{F}}, \lambda_{\text{Sp}} \geq 0$ balance the contributions of temporal, frequency, and spatial edges, respectively, with $\lambda_{\text{T}} + \lambda_{\text{F}} + \lambda_{\text{Sp}} = 1$. Consequently, $W_q(u,v) \in [0,1]$, and the threshold $\eta \in [0,1]$ specifies the minimum edge strength allowed during expansion. We retain only edges satisfying this threshold, forming the graph $\mathcal{G}_q^\eta = (\mathcal{V}, \mathcal{R}_q^\eta)$ with

$$\mathcal{R}_q^\eta = \{(u,v) \in \mathcal{R} \mid W_q(u,v) \geq \eta\}. \tag{13}$$

The multi-hop candidate set is

$$\mathcal{V}_{\text{hop}} = \{v_j \in \mathcal{V} \mid d_{\mathcal{G}_q^\eta}(v_j, \mathcal{V}_{\text{root}}) \leq L_{\text{hop}}\}, \tag{14}$$

where $d_{\mathcal{G}_q^\eta}(v_j, \mathcal{V}_{\text{root}})$ is the shortest hop distance from a node to any root in the thresholded graph; we use $L_{\text{hop}} = 2$ by default. QCER traverses eligible frontier edges in descending order of W_q and collects their endpoint nodes, yielding a coarse temporal–frequency–spatial neighborhood around the query-relevant roots.

For a stream of length T with $N(T) = |\mathcal{V}|$ nodes, collection stops at the dynamic budget

$$K(T) = \min\{K_{\max}, N(T), \max(K_{\min}, \lceil \gamma N(T) \rceil)\}, \tag{15}$$

or when no eligible frontier edge remains. The resulting set stays connected to at least one root, and edge weights determine both candidate eligibility and traversal order. The top-$K(T)$ control uses the same budget but ranks nodes independently using U_{seed}.

Let $\mathcal{N}_q \subseteq \mathcal{V}_{\text{hop}}$ denote the final retrieved set. For each $v_{b,f} \in \mathcal{N}_q$, QCER loads $X_{b,f}$ from the path $\pi_{b,f}$ stored in $m_{b,f}$, encodes the frame, and projects its visual tokens into the MLLM input space. Thus, the graph stores only frame metadata and embeddings, while frame tensors are materialized after retrieval.

4 Experiments

4.1 Experimental Settings

Evaluation Metrics. We evaluate on EventMind and EventChat-Sub; detailed dataset descriptions and split protocols are provided in Appendix **??**.

Following EventFlash, we evaluate generalization and reasoning with four task metrics aligned with protocols used in LLaVA and related MLLM benchmarks (Liu et al., 2023; 2026). More precisely, we use (i) global detailed captioning (GDC) to assess scene-level summarization, (ii) fine-grained question answering (FGQA) to evaluate localized spatial and temporal event details, (iii) multiple-choice question answering (MCQA) to measure instruction following and discriminative reasoning over candidate answers, and (iv) human-action question answering (HAQA) to measure temporal reasoning about human actions. For the open-ended GDC and FGQA tasks, we use GPT-5.6 Luna as an LLM judge, consistent with prior benchmarks. For MCQA and HAQA, we report accuracy based on normalized exact matches to the ground-truth answers. Efficiency is evaluated only for the online inference stage and is reported as generated tokens per second.

Table 1: Comparison of video-based MLLMs and event-based MLLMs on the EventMind and EventChat-Sub datasets.

Method	Params	LLM	Thrpt.	EventMind				EventChat-Sub	
			Token/s	GDC	FGQA	MCQA	HAQA	GDC	FGQA
Qwen2.5-VL	3B	Qwen2.5	–	20.6	41.7	34.6	23.8	34.5	51.2
VideoChat2-Flash	2B	Qwen2.5	–	31.6	38.9	43.6	16.2	36.9	43.8
InternVL2.5	4B	Qwen2.5	–	17.9	37.0	27.3	21.3	28.9	44.6
VideoChat2-Flash	7B	Qwen2.5	–	36.2	41.9	48.2	18.9	53.1	53.6
LLaVA-NeXT-Video	7B	Qwen2.5	–	31.2	44.6	42.7	22.8	46.3	54.8
Qwen2.5-VL	7B	Qwen2.5	–	22.1	43.9	41.8	28.6	41.6	53.2
InternVL2.5	8B	InternLM2.5	–	19.7	40.0	38.2	25.3	42.5	55.6
EventGPT	7B	Vicuna-v1.5	**42.2**	–	–	–	–	71.2	78.2
EventFlash	3B	Qwen2.5	28.5	46.8	61.1	60.0	84.9	71.5	78.6
EventFlash	7B	Qwen2.5	24.0	52.3	64.2	63.1	87.6	74.1	79.5
Dense baseline	3B	Qwen2.5	23.8	46.2	62.5	59.6	86.8	71.1	77.8
TFE-GraphRAG (Ours)	3B	Qwen2.5	34.7	49.7	64.5	64.8	87.8	72.3	79.3
TFE-GraphRAG (Ours)	7B	Qwen2.5	31.6	**55.6**	**67.8**	**67.5**	**89.9**	**76.4**	**80.3**

Implementation Details. We initialize the event encoder with CLIP-ViT-Large-Patch14 (Radford et al., 2021) and use Qwen2.5 as the LLM backbone. A two-layer multilayer perceptron serves as the Event-Language Projector to align the event and semantic spaces. For both the 3B and 7B backbones, we load the Qwen2.5-Instruct checkpoints from Hugging Face[1] and use the LLMs directly without fine-tuning. For throughput evaluation, inference is deployed with Hugging Face Transformers on a single NVIDIA RTX 4090 GPU, following the deployment protocol used by EventFlash (Liu et al., 2026).

4.2 Comparison with State-of-the-Art MLLMs

Table 1 presents experimental results for representative state-of-the-art video-based and event-based MLLMs on EventMind and EventChat-Sub across approximately 3B and 7B scales. The Dense Event Baseline is constructed by removing WEMC and QCER from the 3B TFE-GraphRAG configuration. Overall, our framework improves the corresponding base MLLMs by 3.0%–5.4%. Averaged over the six reported metrics, TFE-GraphRAG surpasses EventFlash by 2.6 points at 3B and 2.8 points at 7B, corresponding to relative gains of 3.8% and 4.0%, respectively.

Because generated-token throughput depends strongly on hardware and deployment, cross-device Token/s values are provided only for reference. Table 1 reports online throughput in generated Token/s, whereas Table 3 additionally reports normalized offline and online runtime in seconds of processing per second of input event stream. In the controlled same-device comparison, TFE-GraphRAG increases online throughput from 23.8 to 34.7 Token/s over the Dense Event Baseline, corresponding to $1.46\times$ the throughput, or a 45.8% increase. For reference, EventGPT reports 35.8 Token/s on an RTX 4090 and 42.2 Token/s on an A100 GPU; the A100 value is a cross-device reference.

4.3 Ablation Study

Component Contribution. Table 2 isolates the contributions of CAS and QCER. CAS alone (A) improves all four task metrics but reduces throughput to 2.4 Token/s, whereas QCER alone (B) achieves the highest throughput of 39.2 Token/s with mixed accuracy changes. Combining both

[1] https://huggingface.co/Qwen.

components yields the best overall accuracy at 34.7 Token/s, improving the Baseline by 3.5, 2.0, 5.2, and 1.0 points on GDC, FGQA, MCQA, and HAQA, respectively.

Table 2: Component contributions on EventMind with the 3B backbone. Throughput is measured in Token/s. C denotes CAS frames and G denotes QCER graph retrieval.

Model	C	G	Throughput	GDC	FGQA	MCQA	HAQA
Baseline	×	×	23.8	46.2	62.5	59.6	86.8
A	✓	×	2.4	46.7	63.6	61.4	87.4
B	×	✓	**39.2**	44.8	62.8	60.1	86.9
Ours	✓	✓	34.7	**49.7**	**64.5**	**64.8**	**87.8**

To isolate the contribution of graph structure, we compare full QCER against variants without the event evidence graph and without each edge type. All variants use the same WEMC memory, retrieval budget, coefficient payloads, and backbone. Without the graph, retrieval reduces to independent semantic ranking; the remaining variants separately remove temporal, frequency, or spatial relations. As shown in Table 3, full QCER improves the no-graph variant by 4.3, 2.7, 4.1, and 4.6 points on GDC, FGQA, MCQA, and HAQA, respectively, while increasing throughput from 32.5 to 34.7 Token/s.

Table 3: Ablation of the event evidence graph and its edge types on EventMind with the 3B backbone. All variants use the same WEMC memory and retrieval budget. Off. and On. denote normalized offline and online time in seconds per second of input.

Variant	Off.	On.	Thrpt.	GDC	FGQA	MCQA	HAQA
w/o graph	31.5	6.9	32.5	45.4	61.8	60.7	83.2
w/o temporal	40.0	6.4	35.2	49.0	64.0	63.2	85.9
w/o frequency	39.8	6.4	35.1	48.6	63.1	63.8	86.4
w/o spatial	32.0	6.2	**36.0**	47.9	63.5	64.1	86.9
Full QCER	40.2	6.5	34.7	**49.7**	**64.5**	**64.8**	**87.8**

Frequency Selection. Table 4 measures whether query-adaptive frequency selection is preferable to a fixed frequency. The three fixed policies restrict retrieval to high-, mid-, or low-frequency nodes, whereas the adaptive policy uses the query-conditioned preference over all three groups. Throughput ranges from 32.8 to 35.9 Token/s, with fixed-mid attaining the highest throughput. Fixed-high achieves the best FGQA score of 64.9. Query-adaptive ties fixed-low for the best GDC score of 49.7 and achieves the best MCQA and HAQA scores of 64.8 and 87.8, respectively, while retaining 34.7 Token/s throughput.

Table 4: Fixed versus query-adaptive frequency selection on EventMind. Throughput is measured in Token/s.

Type	Throughput	GDC	FGQA	MCQA	HAQA
Fixed-high	32.8	47.9	**64.9**	63.2	86.9
Fixed-mid	**35.9**	48.8	63.8	64.1	87.4
Fixed-low	33.6	**49.7**	62.9	63.7	87.2
Query-adaptive	34.7	**49.7**	64.5	**64.8**	**87.8**

4.4 Qualitative Comparison

We present additional qualitative examples in Figure 3. The baseline undercounts the lifts (four versus the ground-truth five), whereas TFE-GraphRAG recovers all five lifts. For the skateboard, the baseline omits the approach, deeper squat, and leftward glide while straightening; TFE-GraphRAG matches the complete ground-truth sequence.

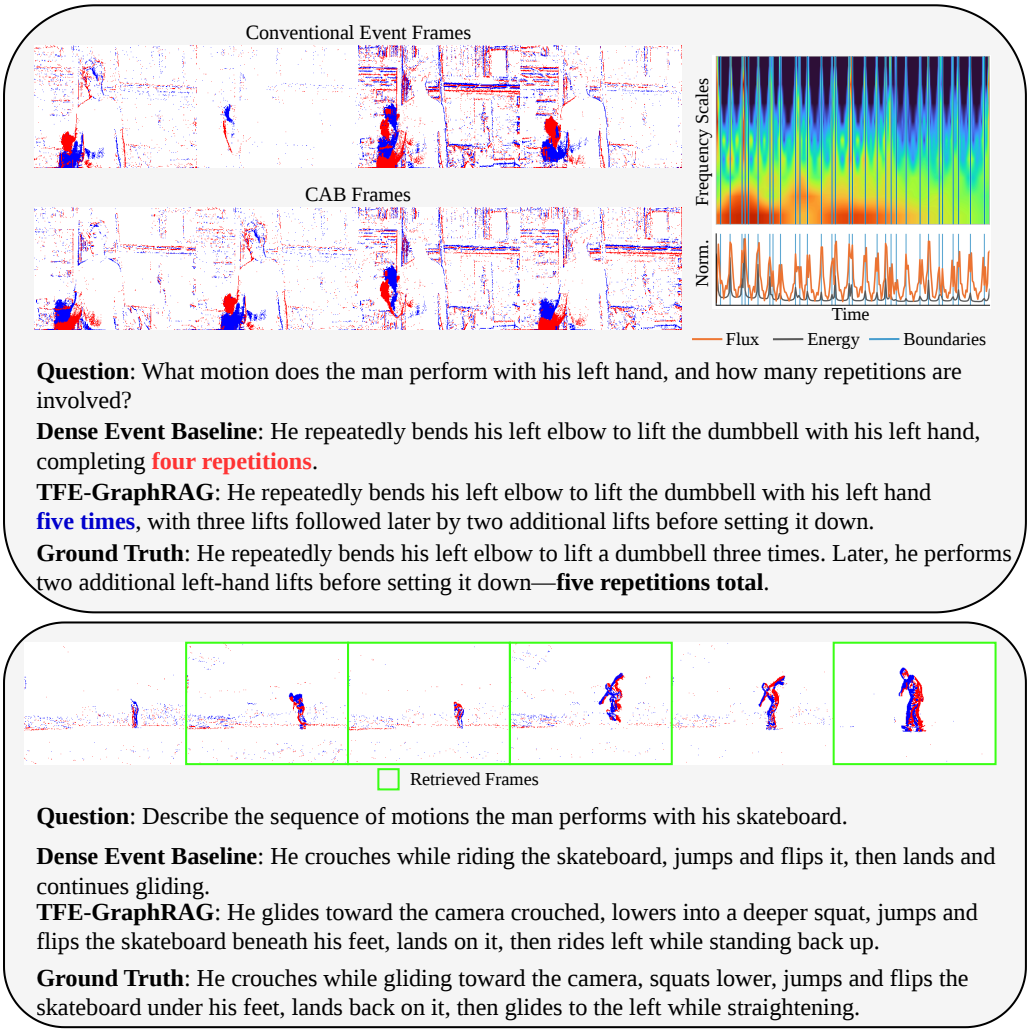

Figure 3: Qualitative examples of adaptive evidence construction and connected retrieval. Four of the 32 bins are displayed in the top example. For visual clarity, the bottom example shows only a subset of the frames retrieved by QCER; inference uses the complete retrieved set.

5 Conclusion

We introduced TFE-GraphRAG, an event-stream MLLM interface that directly addresses the two problems identified in the introduction. CAS mitigates temporal information loss by using patch- and frequency-specific CWT energy changes to adapt the slicing intervals, preserving both rapid transients and sustained motion. The event evidence graph mitigates spatiotemporal context fragmentation by connecting CAS frames through temporal, cross-frequency, and spatial relations, while QCER retrieves connected query-relevant evidence rather than isolated frames. Across the two model scales, TFE-GraphRAG improves over EventFlash by approximately three points on average. Against the same-device Dense Event Baseline, it yields gains of roughly 1–5 points across the EventMind tasks while providing $1.46\times$ the online throughput, a 45.8% increase. The full system requires 40.2 seconds of offline processing and 6.5 seconds of online processing per second of event stream. Compared with independent semantic retrieval, graph-based retrieval improves the EventMind metrics by 2.7–4.6 points, demonstrating the value of preserving relational evidence during retrieval.

AI USE STATEMENT

In this work, we use AI to aid or polish writing and for retrieval and discovery (including finding related work). We also use AI for execution. We use GPT 5.6 Luna to rate video reasoning outputs.

REFERENCES

Inigo Alonso and Ana C. Murillo. EV-SegNet: Semantic segmentation for event-based cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1624–1633, 2019.

R. Wes Baldwin, Ruixu Liu, Mohammed Almatrafi, Vijayan Asari, and Keigo Hirakawa. Time-ordered recent event (TORE) volumes for event cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2519–2532, 2023. doi: 10.1109/TPAMI.2022.3172212.

Jiahang Cao, Mingyuan Sun, Ziqing Wang, Hao Cheng, Qiang Zhang, Shibo Zhou, and Renjing Xu. Spiking neural network as adaptive event stream slicer. In *Advances in Neural Information Processing Systems*, volume 37, pp. 75064–75094, 2024. doi: 10.52202/079017-2389.

Zihan Cheng, Xingyu Pan, Xi Chen, and Shenghua Fan. Wavelet-based feature representation framework for event stream recognition. In *2025 IEEE International Conference on Multimedia and Expo*, pp. 1–6, 2025. doi: 10.1109/ICME59968.2025.11208986.

Honghao Fu, Miao Xu, Yiwei Wang, Dailing Zhang, Jun Liu, and Yujun Cai. VideoStir: Understanding long videos via spatio-temporally structured and intent-aware RAG. *arXiv preprint arXiv:2604.05418*, 2026. URL https://arxiv.org/abs/2604.05418.

Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Joerg 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.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. VTimeLLM: Empower LLM to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14271–14280, 2024.

Xavier Lagorce, Garrick Orchard, Francesco Galluppi, Bertram E. Shi, and Ryad B. Benosman. HOTS: A hierarchy of event-based time-surfaces for pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(7):1346–1359, 2017. doi: 10.1109/TPAMI.2016.2574707.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023.

Pengteng Li, Yunfan Lu, Pinghao Song, Wuyang Li, Huizai Yao, and Hui Xiong. EventVL: Understand event streams via multimodal large language model. *arXiv preprint arXiv:2501.13707*, 2025. URL https://arxiv.org/abs/2501.13707.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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*, pp. 29139–29149, 2025.

Shaoyu Liu, Jianing Li, Guanghui Zhao, Yunjian Zhang, Wen Jiang, Ming Li, and Xiangyang Ji. EventFlash: Towards efficient MLLMs for event-based vision. In *International Conference on Learning Representations*, 2026.

Christoph Posch, Teresa Serrano-Gotarredona, Bernabe Linares-Barranco, and Tobi Delbruck. Retinomorphic event-based vision sensors: Bioinspired cameras with spiking output. *Proceedings of the IEEE*, 102(10):1470–1484, 2014. doi: 10.1109/JPROC.2014.2346153.

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 119336–119360, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Henri Rebecq, Rene Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3857–3866, 2019.

Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. VideoRAG: Retrieval-augmented generation with extreme long-context videos. *arXiv preprint arXiv:2502.01549*, 2025. URL https://arxiv.org/abs/2502.01549.

Xiaoqian Shen, Wenxuan Zhang, Jun Chen, and Mohamed Elhoseiny. Vgent: Graph-based retrieval-reasoning-augmented generation for long video understanding. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Zhipeng Sui, Haiqing Hao, Weihua He, Seng-Hong Lee, and Wenhui Wang. Adaptive spatial-temporal window: Unlocking the potential of event cameras in heterogeneous velocity scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 946–955, 2026.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29118–29128, 2025.

Songyuan Yang, Weijiang Yu, Ziyu Liu, Guijian Tang, Wenjing Yang, Huibin Tan, and Nong Xiao. Graph-to-frame RAG: Visual-space knowledge fusion for training-free and auditable video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33805–33815, 2026.

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*, pp. 22294–22304, 2025.