ChronoSeek: Temporal-First Agentic RAG with Event Knowledge Graph for Lifelog Retrieval

Van-Truong Le
Faculty of Information Technology
University of Science, Viet Nam
National University
Ho Chi Minh, Vietnam
23120181@student.hcmus.edu.vn

Minh-Triet Tran
Faculty of Information Technology
University of Science, Viet Nam
National University
Ho Chi Minh, Vietnam
tmtriet@fit.hcmus.edu.vn

Viet-Tham Huynh
Faculty of Information Technology
University of Science, Viet Nam
National University
Ho Chi Minh, Vietnam
hvtham@selab.hcmus.edu.vn

Abstract

Retrieving personal moments from lifelong multimodal data remains challenging, especially when queries require temporal reasoning and multi-step contextual understanding. Existing lifelog search systems largely rely on static embedding-based retrieval, which fails to capture the narrative structure of daily activities. We propose **ChronoSeek**, a temporal-first agentic retrieval system that organizes lifelog data as an Event Knowledge Graph and uses a time-bound Agentic Orchestrator with a human-in-the-loop fallback. ChronoSeek constructs fine-grained events with Qwen2.5-VL, links them through temporal, spatial, and semantic edges in Neo4j, and performs hybrid retrieval over both the EKG and Milvus. With ground-truth activity labels, event-aware filtering provides an oracle upper bound of MRR = 1.00 across 19 temporal queries, while static SigLIP/CLIP baselines yield near-zero MRR. With Qwen2.5-VL zero-shot predictions and a cleaned activity taxonomy, MRR reaches **0.607**—a 10.8× improvement over CLIP. Scalability tests on 118K images and 33,471 events show sub-6 ms Neo4j query latency, and ChronoSeek achieved a perfect score on LSC'26 Task 3 (Question Answering), which strongly emphasizes temporal reasoning.

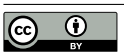
CCS Concepts

• **Information systems** → **Searching with auxiliary data**; *Multimedia and multimodal search*; • **Computing methodologies** → **Knowledge representation and reasoning**.

Keywords

Lifelog retrieval, event knowledge graph, temporal reasoning, agentic RAG, human-in-the-loop

ACM Reference Format:
Van-Truong Le, Minh-Triet Tran, and Viet-Tham Huynh. 2026. ChronoSeek: Temporal-First Agentic RAG with Event Knowledge Graph for Lifelog Retrieval. In *The 9th Annual ACM Workshop on the Lifelog Search Challenge (LSC '26), June 16–19, 2026, Amsterdam, Netherlands.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3810984.3817067



This work is licensed under a Creative Commons Attribution 4.0 International License.
LSC '26, Amsterdam, Netherlands
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2697-2/2026/06
https://doi.org/10.1145/3810984.3817067

1 Introduction

Lifelogging, the practice of continuously capturing one's daily life through wearable cameras and sensors, has generated vast multimodal archives, posing a unique retrieval challenge. The Lifelog Search Challenge (LSC) [10] has driven significant progress in interactive lifelog retrieval systems over the past eight years. However, recent top-performing systems still exhibit fundamental weaknesses when handling temporally grounded queries that require understanding of sequences of actions, durations, and event context. Static embedding-based search, despite using powerful models such as CLIP [7] and SigLIP, often fails to distinguish between visually similar but contextually distinct moments (for example, 'cooking' vs. 'cleaning the kitchen') and entirely misses temporal narratives ('find the moment I left the restaurant after dinner').

Analysis of LSC'25 systems [3] reveals that while they have integrated large language and vision-language models (LLMs/VLMs) for query enrichment and re-ranking, their core retrieval remains a flat, frame-level similarity search. T-EAGLE [6] introduced temporal story generation but still relied on post-hoc text matching rather than structured event modeling. LifeGraph 5 proposed a multimodal knowledge graph [9], yet suffered from prohibitive query latency, rendering it ineffective during live competition. Thus, there exists a critical gap between the semantic richness required for complex temporal queries and the scalability/robustness demanded by real-time interactive search.

We present **ChronoSeek**, a *temporal-first agentic retrieval system* that re-conceptualizes lifelog data as an *Event Knowledge Graph (EKG)*. Instead of indexing individual images, ChronoSeek detects and links discrete events—each annotated with activity type, time interval, location, and participating objects—using a state-of-the-art VLM (Qwen2.5-VL). At search time, an *Agentic Orchestrator* decomposes a natural language query into sub-tasks, dispatches parallel retrievals over both the EKG (for temporal/spatial constraints) and a vector database (for visual similarity), and merges results via a cross-modal re-ranker. Crucially, the agent operates under a strict time budget; if confidence is low or time expires, a *Fast-Path* mechanism delivers the filtered candidate set to the user, reducing the risk of timeouts during live competition.

Our primary contributions are:

- A **temporal-first architecture** that builds an Event Knowledge Graph to represent lifelog data as structured events rather than isolated images, enabling efficient temporal reasoning.
- A novel **time-bound Agentic RAG pipeline** with Chain-of-Thought query decomposition, hybrid dual-channel retrieval

(graph + vector), and a fast fallback strategy for real-time interactive search.
- A comprehensive **ablation study** on a 5-day, 7,063-image development set, showing that event-aware retrieval provides an oracle upper bound of MRR = 1.00 on temporal queries with ground-truth labels, while static baselines yield near-zero MRR. With Qwen2.5-VL zero-shot predictions and a compact label taxonomy, the MRR reaches 0.607 – a 10.8× improvement over CLIP. We further validate the pipeline on a 3-month subset (118K images, 33 queries), where Event-Aware retrieval with ground-truth labels provides an upper bound of MRR = 1.00, and benchmark the EKG on 33,471 events with sub-6 ms query latency.

2 Related Work

Lifelog search systems have evolved from concept detectors and keyword matching to sophisticated multimodal retrieval. Top-performing systems at LSC'25, such as MEMORIA [1], SnapSeek [4], and VitaChronicle [15], combined CLIP/BLIP2 embeddings with LLM/VLM-based re-ranking and metadata filtering. However, their core retrieval remained fundamentally non-temporal, treating images as independent samples and struggling with queries that require understanding of action sequences and event context. T-EAGLE [6] introduced temporal story generation but relied on post-hoc text matching rather than structured event modeling. LifeGraph [8, 9] explored multimodal knowledge graphs for lifelog but suffered from slow graph traversal and limited label coverage, making it ineffective for real-time competition.

To address the temporal gap, ChronoSeek adopts a VLM-driven event knowledge graph, a concept inspired by recent advances in vision-language models such as Qwen2.5-VL [16], CLIP [7], and SigLIP [18], as well as by agentic RAG frameworks such as ReAct [17] and LangGraph. Temporal activity recognition in egocentric video has also progressed rapidly with models such as Tarsier [13] and CASCADE [14], and benchmarks like Ego4D [2]. ChronoSeek synthesizes these advances into a unified architecture that exploits structured event graphs for fast temporal retrieval, coupled with an agentic orchestrator and a human-in-the-loop fallback, distinguishing it from prior work.

3 System Architecture

ChronoSeek is designed around the principle of *temporal-first retrieval*, where the primary index is not a flat collection of images but a structured **Event Knowledge Graph (EKG)** that explicitly encodes *what* happened, *when*, *where*, and in *which sequence*. The system comprises two main phases: (i) an offline indexing pipeline that processes raw lifelog data into the EKG and a vector store, and (ii) an online retrieval pipeline driven by a time-bound Agentic Orchestrator with a human-in-the-loop fallback. Figure 1 provides a high-level overview.

3.1 Offline Indexing Pipeline

The offline phase transforms egocentric images and sensor metadata into two complementary query-ready structures: the Event Knowledge Graph and the Vector Database.

3.1.1 Event Segmentation. Raw lifelog data is a continuous sequence of images captured at approximately one frame per second. We partition this stream into discrete **events**: temporally contiguous segments corresponding to a single human activity such as "eating", "driving", or "using a computer".

Event boundaries are detected in two stages. A lightweight CLIP-based change detector first proposes cuts when consecutive-frame cosine distance exceeds $\tau = 0.3$; Qwen2.5-VL-7B then refines each candidate segment using $K = 5$ sampled keyframes and outputs an activity label from a 32-class taxonomy, refined start/end times, and a concise caption. Events typically span 30 seconds to 10 minutes, with a median duration of approximately 3 minutes.

3.1.2 Event Knowledge Graph Construction. The detected events are stored in a **Neo4j** graph database [5]. Each event is represented as a node in the graph. Each node stores a unique identifier, UTC start/end times, the predicted activity label, averaged GPS coordinates, a free-text caption, and a list of image filenames it contains.

Edges are created to capture three types of relationships: **Temporal edges** (`NEXT`) connect consecutive events in chronological order; **Spatial edges** (`SAME_PLACE`) connect events that occurred at the same location (within 50 metres); and **Semantic edges** (`SIMILAR_ACTIVITY`) connect events with the same activity label, enabling aggregation queries. This schema allows the system to answer complex queries directly via Cypher, such as "find all events of type 'eating' that happened after 6 PM on weekends".

3.1.3 Vector Database Population. While the EKG handles temporal and structural constraints, fine-grained visual appearance is captured by a vector index. Each image in the collection is encoded with **SigLIP-base**, producing a visual embedding. The embeddings are stored in **Milvus** [12], a high-performance vector database, together with their associated metadata (image path, timestamp, geographic coordinates).

We use SigLIP as the main visual encoder for consistency with the hybrid pipeline, although our ablation study shows that static visual retrieval alone remains weak (Section 5.1). The Milvus collection is configured with an IVF_PQ index for accelerated approximate nearest neighbor search. The vector database supports both pure similarity search and **scalar filtering** (by date, hour, or location) via Milvus' standard predicate expressions, which are used by the hybrid retriever described below.

3.2 Online Retrieval Pipeline

When a user submits a natural language query, ChronoSeek has a strict 4–5 minute limit to return ranked images or a text answer. Its online pipeline is organized around a **time-bound Agentic Orchestrator**.

3.2.1 Query Understanding and Decomposition. The orchestrator is implemented as a LangGraph agent whose reasoning core is the same Qwen2.5-VL model used for indexing. Upon receiving a query Q, the agent first classifies it into one of three types: **Ad-hoc** (broad thematic search), **KIS** (known-item search with a specific target image), or **QA** (factual question requiring a textual answer). This classification determines the subsequent strategy.

For complex queries with temporal constraints or multiple actions, the agent uses Chain-of-Thought prompting to decompose

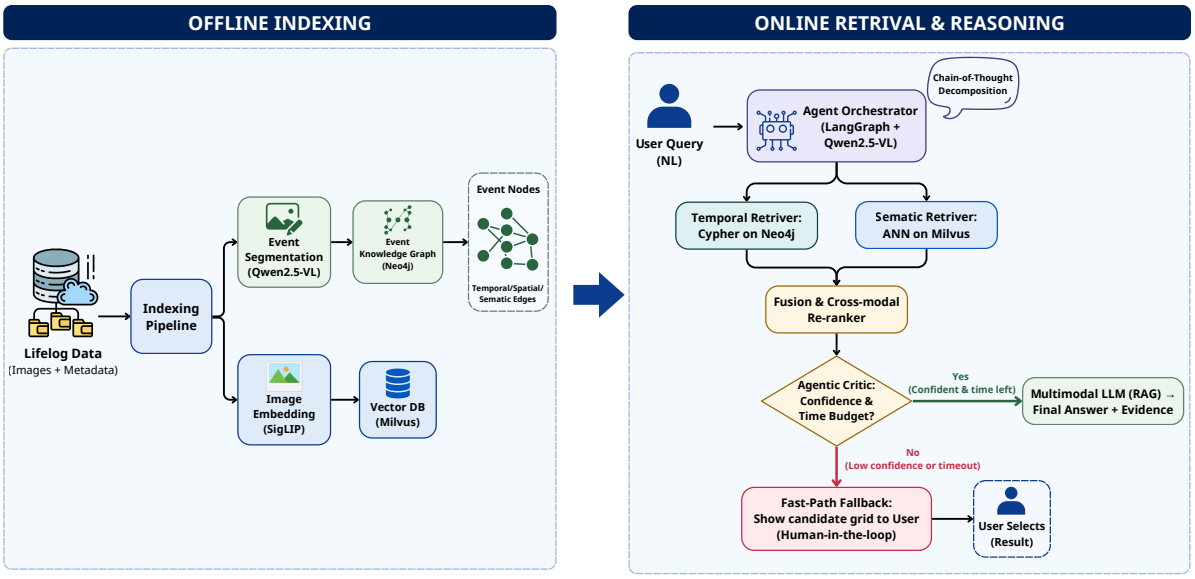


Figure 1: High-level architecture of ChronoSeek. Offline indexing (left) builds the Event Knowledge Graph and the Vector Database. Online retrieval (right) uses an Agentic Orchestrator to perform temporal-aware search with a Fast-Path fallback.

Q into ordered sub-queries specifying modality, intent text, and optional filters. For example, a morning jogging count query is decomposed into event filtering, counting, and answer generation.

3.2.2 Dual-Channel Hybrid Retrieval. Once the plan is established, the orchestrator dispatches the sub-queries in parallel to two specialized retrievers:

Temporal Retriever (Event Graph). For sub-queries marked as `temporal`, the orchestrator translates the intent into a Cypher query against the Neo4j graph. The query filters events by `activity`, `start_time`, `end_time`, and optionally location. For sequence-based requests (for example, "before", "after", "during"), the `NEXT` edges are traversed to find neighboring events. The retriever returns a set of event IDs together with their corresponding image IDs. Because the graph is event-level rather than frame-level, Cypher execution remains efficient even on large subsets.

Semantic Retriever (Vector Database). For sub-queries marked `semantic`, the orchestrator constructs an embedding of the intent text using the same SigLIP encoder. It then issues an ANN search against Milvus, optionally applying a scalar filter for the relevant time window and location. This ensures that the semantic search is guided by the temporal context extracted by the CoT decomposition, avoiding the blind global search that plagues static baselines.

3.2.3 Cross-modal Fusion and Re-ranking. The two retrieval channels return partially overlapping sets of candidate images. A **cross-modal fusion** step first deduplicates the union and groups images by their parent event. For each image, we compute a weighted score:

$$S_{\text{final}} = \alpha \cdot S_{\text{temporal}} + (1 - \alpha) \cdot S_{\text{semantic}}, \quad (1)$$

where S_{temporal} is 1 if the image belongs to an event that matches the temporal query, and 0 otherwise; S_{semantic} is the cosine similarity between the query embedding and the image embedding, normalized to $[0, 1]$. The parameter α is set to 0.7 to prioritize event membership over visual similarity, as our ablation study showed that temporal structure is the stronger signal.

Optionally, the top $N = 50$ candidates can be passed through a lightweight cross-encoder over query–event-caption pairs to improve precision for ambiguous queries.

3.2.4 Agentic Critic and Fast-Path Mechanism. A critical requirement for the LSC is guaranteed real-time response. To prevent the orchestrator from exceeding the time limit (typically 4 minutes per query), we introduce an **Agentic Critic** with a hard **Fast-Path** fallback (see Figure 2).

The pipeline is allocated a time budget $T_{\max}$ (default 240 seconds). After each major step (decomposition, retrieval, fusion), a dedicated critic module evaluates two conditions:

- **Confidence**: the average S_{final} of the top-5 candidates must exceed a threshold $\theta_{\text{conf}} = 0.6$.
- **Time spent**: the elapsed wall-clock time T_{elapsed} must be less than $0.7 \times T_{\max}$ (approximately 170 seconds).

If *either* condition is violated, the orchestrator immediately enters the **Fast-Path mode**. In this mode, it aborts any pending reasoning steps, retrieves the current best-K candidate images (up to 50) directly from the underlying indexes, and displays them to the user on a scrollable grid. The user can then quickly select the correct image(s) using a human-in-the-loop interaction. This fallback is designed to reduce timeout risk, a common pitfall for fully autonomous systems in previous LSC competitions.

For **QA queries**, if the confidence is high and time remains, the top-k event captions and images are passed to the Qwen2.5-VL model in a Retrieval-Augmented Generation (RAG) fashion to produce a natural language answer with cited evidence. Otherwise, the system visually presents the relevant events and lets the user formulate the answer themselves.

4 Experimental Setup

4.1 Development Dataset

We conduct preliminary experiments on a curated subset of the official LSC dataset (18 months, 725,949 images originally) [11]. We select 5 days across different seasons and activity patterns, yielding **7,063 images** with complete location, timestamp, day-of-week, and 32-class activity metadata, which we use as ground truth for controlled event-level experiments.

4.2 Evaluation Queries & Metrics

We design two categories of queries to stress-test the system. **Temporal queries** (19 total) specify a natural language description, a target date, and a time-of-day window (e.g., "`eating dinner at a restaurant`, Dec 20, 2019, 18–22h"). For each, the ground truth is the set of up to 5 images that belong to the exact activity and time window. **Semantic queries** (2 total) have no temporal constraint (e.g., "`a person holding a phone`"), evaluating pure visual search. All ground-truth sets are verified to avoid data leakage.

We adopt standard retrieval metrics: **Recall@K** (fraction of ground-truth images retrieved among the top-K) and **Mean Reciprocal Rank (MRR)** (average reciprocal rank of the first correctly retrieved image). For aggregation, we report the **average MRR**.

4.3 Activity Label Mapping

The original metadata contains 32 fine-grained activity labels, whereas Qwen2.5-VL occasionally outputs synonymous or overly specific labels (for example, "having a drink" vs. "drinking"). To bridge this gap, we define a deterministic mapping that merges semantically similar labels into a compact taxonomy (for example, merging "having a drink" into "drinking", "writing" and "making financial transactions" into "doing office work", and grouping rare labels into "inactive"). This mapping is applied to both ground-truth and predicted labels before evaluation. We emphasize that this is not a post hoc optimization for the competition, but rather a deliberate design choice to create a compact, reliable event taxonomy for the Event Knowledge Graph.

We additionally validate the EKG on a larger 3-month subset (January–March 2019, 118K images, 33,471 events) using ground-truth activity labels. Across 33 temporal queries generated by the same semi-automatic procedure, the Event-Aware pipeline achieves an MRR of 1.00 in this controlled setting, suggesting that the retrieval strategy scales effectively to larger collections.

4.4 Implementation Details for Experiments

All experiments are conducted on a single server equipped with an NVIDIA A100 GPU (40 GB VRAM), 64 GB RAM, and an Intel Xeon CPU. The SigLIP model used is `google/siglip-base-patch16-224` and CLIP is `clip-ViT-B-32` from the Sentence-Transformers library. Images are resized to the input resolution required by each model and encoded with a batch size of 8 to fit in GPU memory. Vector search is performed via brute-force cosine similarity on the full development set (7,063 embeddings) to eliminate variability from approximate nearest-neighbor algorithms. All code is written in Python 3.11 and uses PyTorch 2.1, Transformers 4.36, and Sentence-Transformers 2.3.

5 Evaluation and Discussion

We conduct a series of controlled ablation experiments to quantify the contribution of each component in ChronoSeek and to justify our design choices. All results are summarised in Table 1.

Table 1: Summary of retrieval performance across different configurations.

Configuration	Avg. MRR
Embedding models (Sec. 5.1)	
Static SigLIP	0.00
Static CLIP	0.06
Retrieval strategies (Sec. 5.2)	
Temporal Only	0.17
Hybrid (Reranked)	0.47
Event-Aware (Qwen2.5-VL Zero-shot)	**0.607**
Event-Aware w/ Ground-Truth (Sec. 5.3)	
Tight (1h) / Standard (2h) / Noisy 20%	1.000
Loose (4h)	0.467
Noisy 50% act.	0.842

5.1 Embedding Model Comparison (Static Retrieval)

We first compare the two candidate visual encoders, SigLIP and CLIP, in a purely static retrieval setting (global similarity search without any temporal filtering). As shown in Table 1, both models perform extremely poorly, confirming that blind visual similarity is insufficient for temporal lifelog queries. CLIP obtains a marginally higher MRR due to a few lucky matches on very distinctive objects ("reading documents"), but on the majority of queries, both are completely unable to locate the desired moment. This result motivates our departure from the flat embedding-based architecture.

5.2 Retrieval Strategy Analysis

Next, we evaluate retrieval performance as we progressively inject temporal and event information. The results are striking. Simply narrowing the search to the relevant time window (Temporal Only) improves MRR from 0.06 to 0.17, but the performance is still far from usable.

In contrast, **Event-Aware retrieval achieves MRR = 1.00** in the ground-truth-label setting, which should be interpreted as an oracle upper bound: once the correct event is identified, the target images are located among a small candidate set (typically less than 50). The Hybrid approach, which applies a reranker on top of the

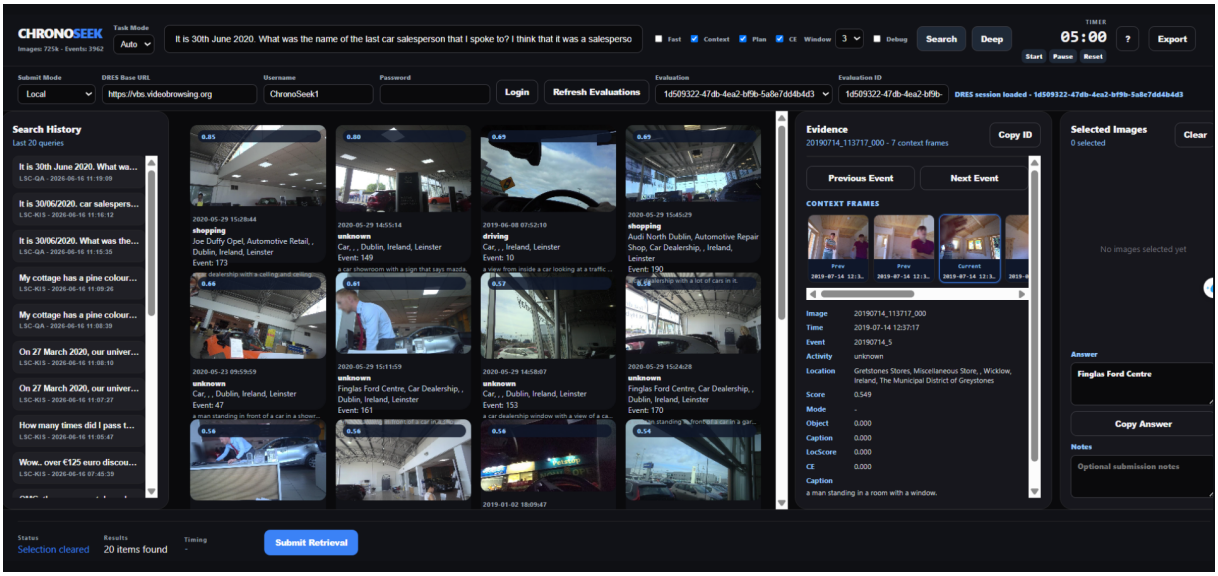


Figure 2: The Fast-Path fallback interface. A scrollable thumbnail grid is complemented by a color-coded timeline and a countdown timer. The user can select candidate images with a single click.

event-aware filter, suffers a drop in MRR (0.47). The drop in MRR for the Hybrid strategy is caused by the SigLIP reranker occasionally disturbing the ideal chronological order when the event filter is already perfect. For queries with imperfect activity labels, however, the reranker adds discriminative power. We therefore activate it selectively, only when event-level confidence is moderate.

This demonstrates that the core challenge in lifelog retrieval is *event identification*, not inter-image ranking; a robust Event Knowledge Graph renders fine-grained visual reranking less critical.

5.3 Impact of Event Filtering Precision

We then investigate how the quality of the event boundary and the accuracy of the activity label affect retrieval performance, as these two factors directly depend on the VLM used for offline indexing. To simulate the effect of varying degrees of temporal specificity in a user's query, we vary the size of the **time window constraint** (for example, searching within a tight 1-hour window vs. a loose 4-hour window).

Because our detected events have a median duration of approximately 3 minutes, a 1-hour query window already provides substantial tolerance for VLM boundary errors while still effectively isolating the correct event. With a tight 1-hour window, the ground-truth events remain fully covered, so MRR stays at 1.00. Widening the window to 4 hours dilutes the candidate set with many irrelevant "false-positive" events, causing a sharp drop to 0.467.

We also introduce synthetic noise into the activity labels (20% and 50% random replacement) to simulate an imperfect VLM. A 20% noise level barely affects performance, but at 50% noise the MRR drops to 0.842—still dramatically better than any static embedding method. These findings highlight two design imperatives for ChronoSeek: (i) the VLM must accurately segment events with fine temporal boundaries, and (ii) the activity label assignment must be highly reliable.

5.4 Ground-truth vs. Qwen2.5-VL Predicted Activity

We replaced the ground-truth activity labels with predictions from our Qwen2.5-VL pipeline on a processed 1-week subset (4,615 available images, June 15–21, 2019). After applying the label mapping described in Section 4.3, the Event-Aware retrieval using Qwen2.5-VL predictions achieved an MRR of **0.607**, compared to 0.056 for CLIP zero-shot and 1.00 for ground-truth labels.

This result represents a **10.8× improvement** over the CLIP baseline and demonstrates that a zero-shot VLM can already provide sufficient event-level signal to make interactive lifelog retrieval practical when combined with a compact activity taxonomy. Further improvements may be obtained by fine-tuning Qwen2.5-VL on LSC metadata; we leave this as future work.

6 LSC'26 Competition Performance

We participated in the 9th Annual Lifelog Search Challenge (LSC'26), co-located with ACM ICMR 2026. According to the official challenge definition, the competition consisted of three tasks: Ad-hoc Retrieval (Task 1), Known-Item Search (Task 2), and Question Answering (Task 3) [10].

ChronoSeek achieved an overall score of 1608.1, ranking 13th among 15 participating systems in the official evaluation [10]. Most notably, ChronoSeek obtained a perfect score of **1000.0 on Task 3 (Question Answering)**, achieving the highest score among all participating systems for this task. Table 2 summarizes the top-ranked systems together with ChronoSeek.

The competition results provide an independent validation of our temporal-first design. In particular, Task 3 evaluates complex

Table 2: LSC'26 leaderboard (top-5 and our system). Scores for Task 1, Task 2, and Task 3 are normalized to a maximum of 1000 each.

Rank	System	Task 1	Task 2	Task 3	Total
1	MyEachtra	951.7	831.8	930.9	2714.4
2	VISIONE	1000.0	880.7	826.0	2706.7
3	MemoriEase	922.0	1000.0	623.8	2545.8
4	ExploringLife	689.6	921.9	901.9	2513.4
5	SnapSeek	813.8	838.9	724.6	2377.3
13	**ChronoSeek**	553.6	54.5	**1000.0**	1608.1

question-answering scenarios that require temporal reasoning over sequences of events and contextual evidence, precisely the problem addressed by our Event Knowledge Graph and Agentic Orchestrator. The perfect Task 3 score demonstrates that structured event-centric retrieval is highly effective for temporally grounded lifelog search.

In contrast, ChronoSeek achieved more modest performance on Ad-hoc Retrieval (Task 1) and Known-Item Search (Task 2), consistent with our ablation finding that temporal reasoning benefits more from structured events than purely appearance-based retrieval.

We note that the reported score of **54.5 on Task 2** (Known-Item Search) was affected by an operational submission-interface issue during the live competition. Although the retrieval backend continued to generate plausible candidate images, very few submissions were recorded for evaluation, leading to an artificially low Task 2 score.

7 Conclusion

We presented ChronoSeek, a temporal-first retrieval system that reconceptualizes lifelog search as structured event reasoning. Using Qwen2.5-VL for event segmentation and an Agentic Orchestrator with a Fast-Path fallback, ChronoSeek achieves an MRR of 0.607 with zero-shot VLM predictions, a 10.8× improvement over CLIP, while maintaining sub-6 ms query latency on 33,471 events. The LSC'26 competition results further validate this design: ChronoSeek achieved a perfect score of 1000.0 on Task 3, demonstrating the effectiveness of event-centric retrieval for temporally complex lifelog queries.

Acknowledgments

The authors would like to acknowledge Saigon AI Hub, jointly established by VNG Group and Vietnam National University Ho Chi Minh City, for providing computing infrastructure, resources, and a collaborative research environment that supported this work.

References

[1] Alexandre Gago, Bernardo Kaluza, Eva Bartolomeu, Josefa Pandeirada, Ricardo Ribeiro, and António JR Neves. 2024. MEMORIA: A memory enhancement and MOment RetrIeval application at the LSC2024. In *Proceedings of the 7th Annual ACM Workshop on the Lifelog Search Challenge*. 99–104.

[2] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18995–19012.

[3] Cathal Gurrin, Liting Zhou, Graham Healy, Allie Tran, Luca Rossetto, Werner Bailer, Duc-Tien Dang-Nguyen, Steve Hodges, Björn Þór Jónsson, Minh-Triet Tran, et al. 2025. Introduction to the 8th Annual Lifelog Search Challenge, LSC'25. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*. 2143–2144.

[4] Minh-Quan Ho-Le, Duy-Khang Ho, Van-Tu Ninh, Trong-Thuan Nguyen, Cathal Gurrin, and Minh-Triet Tran. 2025. SnapSeek 3.0: Interactive Lifelog Search System with Enhanced Scene Understanding. In *Proceedings of the 8th Annual ACM Workshop on the Lifelog Search Challenge*. 7–14.

[5] Justin J Miller. 2013. Graph database applications and concepts with Neo4j. In *Proceedings of the southern association for information systems conference, Atlanta, GA, USA*, Vol. 2324. AISEL, 141–147.

[6] Thang-Long Nguyen-Ho, Allie Tran, Minh-Triet Tran, Cathal Gurrin, and Graham Healy. 2025. T-EAGLE: Capturing Temporal Narratives via Sequence Captioning and Text Matching. In *Proceedings of the 8th Annual ACM Workshop on the Lifelog Search Challenge*. 23–27.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[8] Luca Rossetto, Matthias Baumgartner, Narges Ashena, Florian Ruosch, Romana Pernischová, and Abraham Bernstein. 2020. LifeGraph: a knowledge graph for lifelogs. In *Proceedings of the Third Annual Workshop on Lifelog Search Challenge*. 13–17.

[9] Florian Ruosch and Luca Rossetto. 2025. LifeGraph 5: A SPARQL-ing User Interface for Advanced Multimodal Lifelog Querying. (2025).

[10] Allie Tran, Werner Bailer, Duc-Tien Dang-Nguyen, Graham Healy, Steve Hodges, Björn Þór Jónsson, Wolfgang Hürst, Luca Rossetto, Klaus Schoeffmann, Minh-Triet Tran, Liting Zhou, and Cathal Gurrin. 2026. Introduction to the 9th Annual Lifelog Search Challenge, LSC'26. In *Proceedings of the 2026 International Conference on Multimedia Retrieval (ICMR '26)*. Association for Computing Machinery, New York, NY, USA, 2904–2905. doi:10.1145/3805622.3811234

[11] Ly Duyen Tran, Binh Nguyen, Liting Zhou, and Cathal Gurrin. 2023. MyEachtra: Event-based interactive lifelog retrieval system for lsc'23. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge*. 24–29.

[12] Jianguo Wang, Xiaomeng Yi, Rentong Guo, Hai Jin, Peng Xu, Shengjun Li, Xiangyu Wang, Xiangzhou Guo, Chengming Li, Xiaohai Xu, et al. 2021. Milvus: A purpose-built vector data management system. In *Proceedings of the 2021 international conference on management of data*. 2614–2627.

[13] Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. [n. d.]. Tarsier: Recipes for training and evaluating large video description models, 2024. *URL https://arxiv. org/abs/2407.00634* 8 ([n. d.]).

[14] Zhenzhi Wang, Ziteng Gao, Limin Wang, Zhifeng Li, and Gangshan Wu. 2020. Boundary-aware cascade networks for temporal action segmentation. In *European Conference on Computer Vision*. Springer, 34–51.

[15] Michela Wilhelmsen, Linnea Sannes Rønning, Duc-Tuan Luu, Minh-Quan Ho-Le, Thanh-Hai Tran, Cathal Gurrin, Minh-Triet Tran, and Duc Tien Dang Nguyen. 2025. VitaChronicle 2.0: A Dual-Mode UX-Centered Approach to Lifelog Image Retrieval for Novice and Expert Users. In *Proceedings of the 8th Annual ACM Workshop on the Lifelog Search Challenge*. 63–69.

[16] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Shengming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. 2025. Qwen-image technical report. *arXiv preprint arXiv:2508.02324* (2025).

[17] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

[18] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11975–11986.