ForeSea: AI Forensic Search with Multi-modal Queries for Video Surveillance

Hyojin Park[1*], Yi Li[1*],
Janghoon Cho[1†], Sungha Choi[2†‡], Jungsoo Lee[1†],
Taotao Jing[1], Shuai Zhang[1], Munawar Hayat[1], Dashan Gao[1],
Ning Bi[1], and Fatih Porikli[1]

[1] Qualcomm AI Research, San Diego, CA, USA
[2] Kyunghee University, Gyeonggi, South Korea

Abstract. Despite decades of work, surveillance still struggles in searching and reasoning about specific targets across long, multi-camera videos. Existing methods—tracking, retrieval, and video LLMs—require heavy manual filtering, capture only shallow attributes, and fail at temporal understanding. Prior benchmarks are also limited to basic retrieval and question answering, without addressing real-world challenges that often involve multimodal queries and temporal grounding (e.g., "When did this person join the fight?" with the person's image). To address this gap, we introduce `ForeSeaQA`, a new benchmark specifically designed for video QA with image-and-text queries and timestamped annotations of key events. The dataset consists of long-horizon surveillance footage paired with diverse multimodal questions, enabling systematic evaluation of retrieval, temporal grounding, and multimodal reasoning in realistic forensic conditions. Not limited to this benchmark, we propose `ForeSea`, an AI forensic search system with a 3-stage, plug-and-play pipeline. (1) A tracking module filters irrelevant footage; (2) a multimodal embedding module indexes the remaining clips; and (3) during inference, the system retrieves top-K candidate clips for a video LLM to answer queries and localize events. On `ForeSeaQA` benchmark, `ForeSea` improves accuracy by 3.1 points and temporal IoU by 10.1 points over prior retrieval-augmented baselines. To our knowledge, `ForeSeaQA` is the first benchmark to support complex multimodal queries with precise temporal grounding, and `ForeSea` is the first VideoRAG system built to excel in this setting.

1 Introduction

Recent large multimodal models (LMMs) have made rapid progress in long-form video analysis, driven by advances in general video understanding [7, 47, 51], temporal grounding [34, 37], and complex reasoning [10, 11]. These skills are crucial for applications to video surveillance analysis [23, 33, 45], which requires

[*] Equal contribution as first authors.
[†] Equal contribution as second authors.
[‡] This work was done while the author was at Qualcomm.

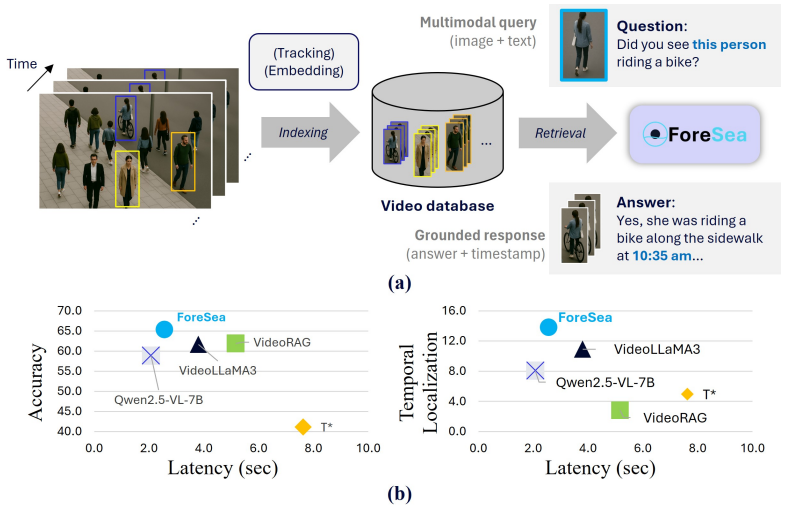


Fig. 1: AI Forensic Search with `ForeSea`. Our proposed framework for long surveillance videos supports complex *multimodal queries* (e.g., a reference image combined with a text question) and leverages a person-centric multimodal database to efficiently retrieve and generate *temporally grounded* answers.

finding specific people, objects, or events of interest across hours or even days of videos captured by multiple cameras.

Existing surveillance systems have traditionally relied on object detection and tracking pipelines [4, 27, 50, 56] to process large-scale video data. While this enables basic analytics such as counting and searching, it struggles with recognizing complex activities, detecting various anomalies, and achieving a holistic understanding of long videos. Each of these tasks often involves substantial human effort, including reviewing retrieved footage, and gathering visual evidence to draw conclusions.

To mitigate this manual effort, recent approaches use CLIP-based models, vision-language models (VLMs) and retrieval-augmented generation (RAG) [6, 21, 24, 29, 30, 36]. However, they still face three key limitations: (1) it only supports text queries, cannot handle multimodal queries; (2) the provided answers are lack of temporal evidence reasoning; and (3) VLM model's accuracy and length of input are constrained in long video. These limitations highlight the need for practical surveillance tasks, such as answering *multimodal* queries with *temporally grounded* evidence (Fig. 1). For example:

Q: "Did you see *this person* riding a bike?" + an image of the individual
A: "Yes, riding a bike *at 10:35 am* along the sidewalk." + a trimmed video clip.

While such tasks are essential for real-world scenarios, they remain absent from existing research. To bridge this gap, we introduce `ForeSeaQA`, the first benchmark for multimodal, temporally grounded video question answering in surveillance. `ForeSeaQA` is built from UCF-Crime [33] videos using a semi-automated

data engine that extracts person entities from dense captions [45] and visually grounds them via a multimodal LLM. The AI then generates QA pairs with precise temporal annotations across six subtasks: *search*, *activity*, *event*, *temporal*, *counting*, and *anomaly*. Crucially, person-specific questions use *multimodal* queries—a reference image of the individual with the question text—to reflect real forensic workflows. All QA pairs are manually verified for validity, unambiguity, and temporal-grounding correctness. To our knowledge, `ForeSeaQA` is the first benchmark to jointly evaluate multiple-choice accuracy and temporal localization under both text-only and multimodal query conditions in surveillance.

We further present `ForeSea`, a simple yet strong multimodal RAG framework combining three off-the-shelf components: (i) a person tracker that segments long videos into person-centric clips, drastically reducing the search space; (ii) a multimodal encoder that indexes clips in a unified image-text embedding space for text and image-text retrieval; and (iii) a videoLMM that reasons over the top-K retrieved clips to produce a temporally grounded answer. Despite its simplicity, `ForeSea` achieves strong performance on `ForeSeaQA` and generalizes to open-domain long-video benchmarks, showing that person-centric retrieval is a powerful inductive bias for surveillance understanding.

We evaluate `ForeSea` on `ForeSeaQA` against off-the-shelf video LMMs and retrieval-augmented baselines. `ForeSea` achieves the best overall accuracy (66.0%) and temporal localization IoU (13.6%), and ranks first on `ForeSeaQA`$^{\texttt{MM}}$ accuracy (65.4%), with the largest gains on the *search* task where person-centric retrieval is most critical. It further generalizes beyond surveillance to open-domain long-video benchmarks, matching or exceeding state-of-the-art methods while using only half as many input frames. `ForeSea` also achieves lower end-to-end latency than all retrieval-augmented baselines (2.6 s vs. 5.2–7.6 s) and VideoLLaMA3 (3.8 s), showing that person-centric retrieval reduces the Video LMM frame budget without sacrificing accuracy.

Our main contributions are as follows. First, we introduce `ForeSeaQA`, the first benchmark for multimodal, temporally grounded video QA in the surveillance domain, covering six subtasks with joint multiple-choice accuracy and temporal localization evaluation under both text-only and multimodal queries. Second, we present `ForeSea`, a simple yet strong Video-RAG baseline that combines off-the-shelf person tracking, multimodal embedding, and a VideoLMM into a unified pipeline for forensic search. Finally, through comprehensive experiments, we show that `ForeSea` outperforms standard Video LMMs and retrieval-augmented baselines on `ForeSeaQA`, generalizes to open-domain long video benchmarks with competitive performance at half the frame budget, and achieves substantially lower retrieval latency than prior RAG approaches.

2 Related Work

Video LMMs. Recent LMMs advance video-language reasoning through two main directions: (1) modality integration, where models like Video-LLaVA [22] and LLaVA-NeXT-Interleave [20] align or interleave visual tokens with text for

multi-frame understanding; and (2) scalability, with VideoLLaMA3 [47] applying token compression for long videos, while InternVL [9] and Qwen2.5-VL [3] leverage large-scale multimodal data and powerful language backbones. Despite these advances, most Video LMMs process the full video end-to-end without external knowledge grounding, which limits performance on long-horizon QA tasks where relevant evidence is sparse.

Retrieval-Augmented Video Understanding. VideoRAG systems combine retrieval from large-scale video corpora with generative models to support long-form video QA. Recent advances include visually-aligned retrieval, graph-based grounding, memory-enhanced retrieval, and adaptive temporal search [15, 25, 26, 29, 30, 41, 43]. In the surveillance domain, video anomaly detection (VAD) methods have adopted language-guided and retrieval-augmented techniques for identifying rare events, including training-free LLM-based scoring, spatiotemporal graph reasoning, and verbalized learning [31, 42, 46, 49]. However, existing VideoRAG systems are designed for general-purpose QA and lack fine-grained temporal localization, while VAD methods target classification or anomaly scoring rather than interactive, multimodal question answering.

Multimodal Retrieval. While cross-modal retrieval focuses on single-modality mappings like image-to-text (e.g., CLIP [28]), multimodal retrieval enables flexible searches across mixed modality pairs [18,53]. Systems such as VISTA [53] and GCL [18] allow queries and targets to include images, text, or both, supporting unified retrieval across heterogeneous inputs. Despite this flexibility, multimodal retrieval remains underexplored for forensic search, where combining image and text queries is crucial for identifying specific individuals.

Benchmarks. General-purpose long video benchmarks, such as InfiniBench [1], LoVR [5], and LongerVideos [29], support long-form retrieval but lack detailed temporal annotations and multimodal query support. Domain-specific benchmarks like TUMTraffic-VideoQA [55], SurveillanceVQA-589K [23] and SmartHome-Bench [52] address traffic, surveillance, and smart home scenarios but restrict queries to a single modality. Event-focused datasets like MomentSeeker [44] emphasize temporal retrieval but target single events rather than complex forensic contexts. `ForeSeaQA` is the first benchmark to jointly evaluate multiple-choice accuracy and temporal localization under both text-only and multimodal query conditions in the surveillance domain.

3 `ForeSeaQA`: Benchmarking Grounded Multimodal Video Understanding

We construct the `ForeSeaQA` benchmark to evaluate the ability of LMMs to understand long videos, ground people and moments of interest, and answer questions based on the retrieved evidence.

3.1 Benchmark Design

The benchmark differs from existing long video benchmarks by introducing two unique challenges to the models.

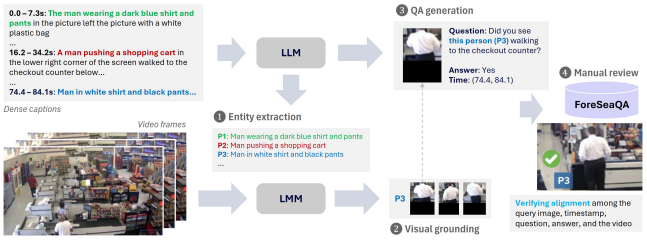


Fig. 2: `ForeSeaQA` Data Engine. We use text-only and multimodal LLMs to ❶ extract person entities from dense video captions, ❷ visually ground each entity to create query image crops, and ❸ generate multimodal QA pairs with timestamps. All generated QA samples and query images are ❹ reviewed by human workers for correctness.

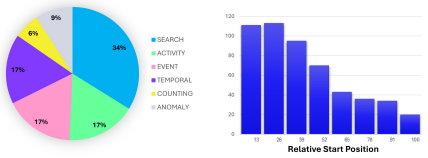
Joint answer and localization. We augment each question-answer pair with time ranges of grounded evidence that supports the answer, and require models to jointly output its answer with the associated timestamps. Specifically, we construct the dataset as $\mathcal{D} = \{(V, Q, A, T)\}$, where T can be one or multiple intervals $T = \{(T_s, T_e)\}$ that contain sufficient and necessary information from video V to predict the correct answer A of question Q. While such time annotations are used in some existing benchmarks (e.g., Charades-STA [13], VideoSIAH [40]) benchmarks, they are often limited to a single interval or a list of non-exhaustive keyframes per question, and usually do not evaluate localization and question answering tasks jointly.

Multimodal queries. In addition to text-only questions, `ForeSeaQA` includes *multimodal* queries with supplementary images to the question. Concretely, each multimodal query is represented as $Q = (Q_I, Q_T)$ where Q_I is an image and Q_T is a question that *refers to* the query image (e.g. "When did *this person* enter the building?"). This mirrors practical scenarios in surveillance analysis, where a snapshot of a person of interest is provided as reference to enable tasks such as identifying when and where the individual appears, or what activities they participate in; answering such questions require LMMs to simultaneously understand the video frames, the reference image and the question interleaved in the same multimodal input sequence, a capability rarely examined in prior video benchmarks.

3.2 Data Engine

We use videos from the UCF-Crime dataset [33] and a semi-automated data engine to generate *temporally grounded* and *multimodal* video QA from dense captions, as illustrated in Figure 2. The engine has 4 stages:
❶ **Entity extraction:** A text-only LLM[3] parses dense UCA [45] captions to extract human entity references (e.g., "man in white shirt"). Multiple references to the same individual are grouped, creating a list of timestamps per person.

[3] Qwen3-32B [39] for QA generation and Qwen2.5-VL-32B [3] for spatial grounding.



(a) (b)

Statistics Length (sec)	
Min	77.35
Max	2112.88
Mean	352.94
Median	262.83
Std	374.60
25th	142.95
75th	399.16

(c)

Benchmark	Tasks	T_ann	MMq
Comprehensive			
LongVid [38]	MC	✗	✗
LVBench [35]	MC	✓	✗
Vid-MME [12]	MC	✗	✗
Temporal retrieval			
ICQ-High [48]	TG	✓	✓
MSeeker [44]	TG	✓	✓
Surveillance domain			
TUMT-VQA [55]	MC, STG	✓	✗
SVQA-589K [23]	OE	✓	✗
ForeSeaQA (ours)	MC, TG	✓	✓

(d)

Fig. 3: Statistics of ForeSeaQA benchmark. (a) Task distribution by question. (b) Relative start position of ground-truth time ranges. (c) Statistics of video duration. (d) Comparison of benchmarks. Tasks: MC=multiple-choice, OE=open-ended, TG=temporal grounding, STG=spatiotemporal grounding. T_ann= Temporal annotation, MMq =Multimodal query.

❷ **Visual grounding:** We use a LMM to ground the extracted entities. For each timestamp from ❶, we sample 8 frames uniformly within the annotated timestamp and ask the model to predict bounding boxes for the referred person. We then crop the bounding boxes and prompt the LMM again to verify the person's presence to prevent hallucinated coordinates. The crops of person entities are used as query images in multimodal questions of ForeSeaQA.

❸ **Grounded QA generation:** We then use the text LLM to generate candidate QA pairs from the captions.[4] ForeSeaQA includes questions from 6 subtasks: *search* (SE), *activity* (AC), *event* (EV), *temporal* (TM), *counting* (CT), and *anomaly* (AN). Among these, search, activity, event and temporal questions are *person-specific* and are generated for each person entity; counting and anomaly questions are *global* and generated for the entire video. For each answer, the LLM assigns temporal groundings by selecting time ranges from the timestamp lists obtained in ❶. To create multimodal questions in person-specific tasks, we rephrase the question to refer indirectly to the grounded entity images from ❷ (e.g., using "the person in the photo" instead of "the man in the white shirt").

❹ **Manual verification:** We manually validate all generated QA pairs by 2 stages. 1) Human reviewers verified the semantic alignment of all QA pairs against the video, aggressively removing 50% of the initial generations. 2) We used VideoLLM[5] with UCA captions to detect misalignment again, and then correct every misaligned QA pair, query image and timestamp manually.

3.3 Benchmark Details

Following the procedure described in Section 3.2, we construct the final ForeSeaQA benchmark, which comprises 1,041 curated questions. Figure 3 summarizes key dataset statistics, including the subtask distribution (Figure 3a), the relative starting positions of annotated temporal windows (Figure 3b), and video-length statistics (Figure 3c). The benchmark spans a wide range of video durations and

[4] Generation prompts per question type are provided in the supplemental material.
[5] We employed Qwen3-VL-32B [2] to verify between video and generated QA pairs.

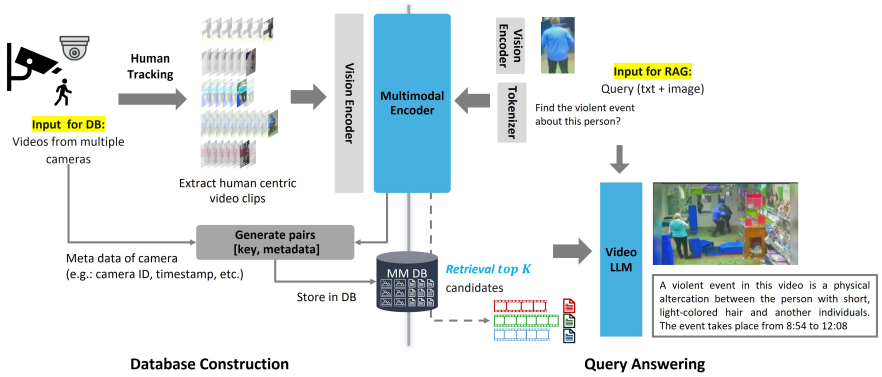


Fig. 4: Overview of `ForeSea` Pipeline. `ForeSea` consists of two main components: (1) Video Database Construction—a multimodal encoder embeds short video clips from the human tracking module and pairs them with metadata; (2) Query Answering—retrieves candidate videos from the database using a multimodal query and generates answers based on the retrieved content

temporal intervals. The starting points of the annotated time ranges vary substantially across questions, demonstrating that temporal grounding in `ForeSeaQA` cannot be solved by heuristics that focus only on early or late portions of the video. While the benchmark places particular emphasis on *search* questions—reflecting their role as a foundation for more advanced temporal reasoning tasks—it also provides balanced coverage of *activity*, *event*, *temporal*, and global tasks such as *counting* and *anomaly detection*. This diversity ensures that models are evaluated across a broad spectrum of forensic video understanding capabilities.

4 Method

We present our `ForeSea`, a novel videoRAG framework designed for multimodal queries. In Sec. 4.1, we describe the overall system architecture about how we build the searchable database, and how our model provides answers for multimodal surveillance queries. In Sec. 4.2, we describe the multimodal encoder in detail. We explain how it encodes visual and textual inputs into a unified embedding space, how these embeddings are stored in the database, and how they are later used during retrieval. Finally, in Sec. 4.3, we explain how the VideoLLM stage answers user queries.

4.1 Overall Architecture

The overall architecture of the proposed system is illustrated in Figure 4. The pipeline consists of two stages: (i) video database construction and (ii) query answering with VideoLMM reasoning.

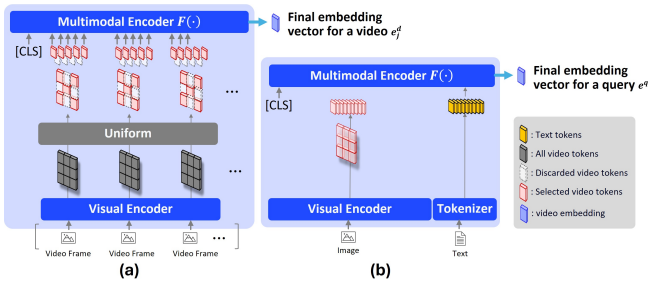


Fig. 5: Multimodal encoder produces (a) a video embedding from multiple frames and (b) a query embedding from text or image-text inputs

Video Database Construction: The system begins by collecting raw video recordings D from multiple cameras. A human tracking module processes these videos to extract only relevant frames, and D is segmented into short clips according to the tracking results. Each segment is then cropped using the corresponding bounding box coordinates to produce human-centric video clips $C = \{c_1, \ldots, c_j\}$. Each clip c_j is fed into the multimodal encoder (detailed in 4.2) to generate a database embedding vector $\mathbf{e}_j^d$. This vector $\mathbf{e}_j^d$, which captures the semantic content of the clip, is stored in a multimodal database together with relevant metadata[6] to enable efficient retrieval.

Query Answering. The system supports various query formats, including text-only queries (q_t) and image–text queries (q_{it}). Given a query, the same multimodal encoder is used to generate a unified query embedding $\mathbf{e}^q$. This vector is matched against the database to retrieve the top-K candidate embeddings $\{\mathbf{e}_j^d\}$. The corresponding top$K$ candidate clips are then concatenated and provided as input to a VideoLMM, along with the original query and augmented information (such as bounding box coordinates), to produce a summary of key events and a temporally grounded answer with linked visual evidence.

4.2 Multimodal Embedding

We build both the retrieval index and the query embeddings using a publicly available multimodal encoder introduced in [18, 53], as shown in Figure 5.

Video embedding: As shown in 5 (a), for each clip c_j from tracking module, we obtain frames $C_j = \{f_{j,k}\}_{k=1}^{m_j}$ and compute frame-level visual tokens $\mathbf{x}_j^d = \{\mathbf{x}_{j,1}^d, \ldots, \mathbf{x}_{j,m_j}^d\}$ with the visual encoder. Here, m_j denotes the total number of frames for the j-th clip. To ensure a consistent number of input tokens for the MMEnc (equivalent to that of a single image input), we apply uniform sampling $S(\cdot)$[7]. We feed the sampled tokens to MMEnc and use the [CLS] output as the

[6] We use camera ID, timestamp, and bounding box coordinates.

[7] The sampling rate adapts to the length of video clips, so the resulting number of tokens matches that of a single-image input. Because the human tracking clips have variable lengths, and uniform sampling provides a fixed-size representation.

database vector $\mathbf{e}_i^d \in \mathbb{R}^p$:

$$\mathbf{e}_j^d = \text{MMEnc}\big([\mathbf{x}_{CLS}, S(\mathbf{x}_j^d)]\big) \tag{1}$$

Query embedding: We use the same MMEnc for all query formats: text (q_t) and image+text(q_{it}) as shown in 5 (b). The text query is tokenized into x_t^q, and an image is encoded by the visual encoder into x_i^q. For a text-only query, x_t^q serves as the input x^q to the MMEnc. For an image-text query, the visual features x_i^q are passed through a projection layer to match the dimension of x_t^q. Both sets of features are then concatenated to form the final input x^q. In all cases, the [CLS] output gives the query vector $e^q \in \mathbb{R}^p$:

$$\mathbf{e}^q = \text{MMEnc}\left([\mathbf{x}_{CLS}, \mathbf{x}^q]\right), \quad \text{where } \mathbf{x}^q \in \{\mathbf{x}_t^q,\ [\mathbf{x}_i^q;\ \mathbf{x}_t^q]\} \tag{2}$$

Most existing approaches perform retrieval in a text-only space, converting all modalities (video frames, ASR transcripts, etc.) into text. This "unimodal projection" inherently leads to information loss. In contrast, ForeSea performs retrieval directly within a unified multimodal embedding space. This approach not only avoids information loss and yields superior accuracy, but it also ensures that semantically relevant instances are retrieved regardless of the query modality, enabling a truly flexible and scalable multimodal search.

4.3 Response Generation from Retrieval Results

Following the retrieval stage, we obtain a set of top-K candidate video clips, each associated with precise spatio-temporal metadata: a start and end timestamp (T_s, T_e) and bounding box coordinates (*bbox*). To prepare input for the videoLMM, we extract frames from each candidate clip at source resolution and draw *bbox* on every frame. This augmentation explicitly directs the model's attention to the people. We guide the VideoLMM's output by providing a system prompt engineered to solicit two key pieces of information: (1) a concise summary of the events occurring within the spatio-temporal window, and (2) a list of precise timestamps for any key events observed.

5 Experiments

In this section, we evaluate a range of existing Video LMMs and retrieval-augmented baselines on `ForeSeaQA`, and present `ForeSea` as a strong baseline for multimodal forensic search. We further demonstrate that `ForeSea` generalizes to open-domain long video benchmarks. Sec. 5.2 presents main results on `ForeSeaQA` under both text-only and multimodal query conditions. Sec. 5.3 and 5.4 ablate the key design choices of `ForeSea`. Sec. 5.5 compares efficiency across methods. Sec. 5.6 evaluates `ForeSea` on VideoMME and MLVU to assess generalization beyond the surveillance domain.

Table 1: Performance comparison on ForeSeaQA. ForeSeaQA^{MM} and ForeSeaQA^{Text} denote multimodal (image+text) and text-only query; ForeSeaQA reports their average.

Model	Params	ForeSeaQA^{MM}		ForeSeaQA^{Text}		ForeSeaQA	
		Acc	IoU	Acc	IoU	Acc	IoU
Video LMMs							
LLaVA-OneVision [19]	7B	56.1	10.4	58.5	7.7	57.3	9.0
GLM-4.1V-Thinking [14]	9B	57.4	10.0	55.1	8.4	56.2	9.2
InternVL3 [57]	2B	38.3	9.8	34.1	7.1	36.2	8.4
	8B	61.3	10.2	63.4	9.9	62.3	10.0
	9B	<u>62.3</u>	11.5	62.7	8.8	62.5	10.2
Qwen2.5-VL [3]	7B	58.9	8.1	59.0	7.5	58.9	7.8
	72B	60.0	**15.3**	61.4	10.1	60.7	12.7
VideoLLaMA3 [47]	7B	61.6	10.9	**67.7**	**15.5**	<u>64.6</u>	<u>13.2</u>
Retrieval-augmented							
VideoRAG [25]	7B	61.9	2.8	63.8	4.3	62.9	3.5
T* [41]	7B	41.1	4.9	48.4	4.2	44.8	4.6
ForeSea **(Ours)**	7B	**65.4**	<u>13.8</u>	<u>66.7</u>	<u>13.3</u>	**66.0**	**13.6**

5.1 Experimental Setup

Evaluation Protocols and Metrics. All evaluations on ForeSeaQA are conducted under two query conditions: *text-only* (ForeSeaQA^{Text}) and *multimodal* image+text (ForeSeaQA^{MM}), as described in Sec. 3. We report *accuracy* (percentage of correctly answered multiple-choice questions) and temporal localization *IoU* (intersection-over-union between the predicted and ground-truth time intervals, averaged over all questions) as the two primary metrics.

Models. We evaluate a diverse set of Video LMMs and retrieval-augmented baselines on ForeSeaQA. For Video LMMs, we include LLaVA-OneVision [19], GLM-4.1V-Thinking [14], InternVL3 [57], Qwen2.5-VL [3], and VideoLLaMA3 [47], spanning model sizes from 2B to 72B parameters. For retrieval-augmented baselines, we include VideoRAG [25] and T* [41]. We also evaluate our proposed ForeSea (Sec. 4).

Implementation Details. ForeSea uses ByteTrack [50] with a YOLO-based [16] detector to segment long videos into person-centric clips, which are indexed using a GCL-trained [18] multimodal encoder following VISTA [53]. During retrieval, it selects the top-K ($K = 3$) most relevant clips and passes them to VideoLLaMA3 [47] for answer generation.

5.2 Results on ForeSeaQA

Main results. We evaluate all models on ForeSeaQA^{Text} and ForeSeaQA^{MM} and calculate their average as the final ForeSeaQA scores; results are reported in Table 1. We highlight three key observations on the benchmark:

Temporal localization is the primary challenge. Despite achieving reasonable multiple-choice accuracy, all Video LMMs produce low temporal localization IoU (7–16%), indicating that correct answers are often inferred from

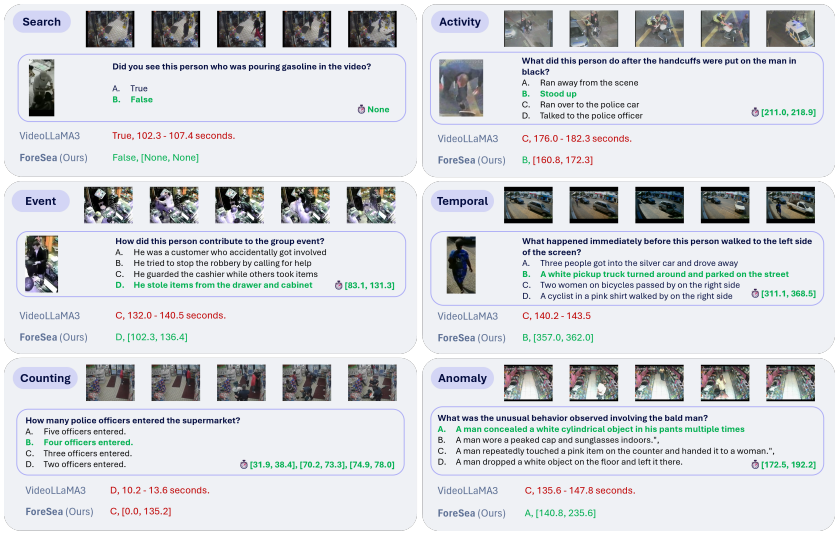


Fig. 6: Qualitative examples of ForeSea and VideoLLaMA3 on ForeSeaQA. Ground-truth answers are highlighted in green. Model answers are highlighted in green if correct (multiple-choice) or have nonzero IoU (temporal grounding), and red if wrong.

global video context rather than grounded evidence. Retrieval-augmented baselines (VideoRAG, T*) fare even worse on IoU (2.8–4.9%), despite comparable or lower accuracy—suggesting that their retrieval strategies do not produce temporally precise evidence. In contrast, ForeSea achieves substantially higher IoU (13.6%), demonstrating that person-centric retrieval is a strong inductive bias for temporal grounding in surveillance videos.

Multimodal queries expose a gap in existing Video LMMs. $\texttt{ForeSeaQA}^{\texttt{MM}}$ is consistently harder than $\texttt{ForeSeaQA}^{\texttt{Text}}$ for most models, with accuracy dropping by up to 6 points (e.g., VideoLLaMA3: 67.7%→61.6%). This suggests that current Video LMMs struggle to jointly reason over a reference image and a long video—a capability central to forensic search. ForeSea is more robust to this shift: it maintains accuracy above 65% on both $\texttt{ForeSeaQA}^{\texttt{Text}}$ (66.7%) and $\texttt{ForeSeaQA}^{\texttt{MM}}$ (65.4%), while no other method does.

Accuracy–localization tradeoff. ForeSea achieves the best overall accuracy (66.0%) and IoU (13.6%) among all retrieval-augmented methods, and ranks first on $\texttt{ForeSeaQA}^{\texttt{MM}}$ accuracy (65.4%) across all evaluated models. Notably, ForeSea outperforms all Video LMMs on $\texttt{ForeSeaQA}^{\texttt{MM}}$ accuracy while using only 7B parameters, demonstrating that person-centric retrieval provides a meaningful advantage over dense video processing for multimodal forensic queries.

Qualitative examples. Figure 6 shows a qualitative comparison between ForeSea and VideoLLaMA3 on samples of different tasks of ForeSeaQA. In *event*, *temporal* and *anomaly* examples, ForeSea correctly identifies the time intervals containing the relevant information and answers correctly, while VideoLLaMA3 fails to

Table 2: Ablation study on ForeSeaQA^{MM}. The highlighted row indicates the default configuration used in the main results. * denotes top K retrieval of *global frames* rather than person-centric clips.

Model	Setup				Multi-choice Acc. (%)					Temporal Loc. IoU (%)				
	Crop	Overlay	Coords	Top K	Search	Act.	Event	Temp.	Avg	Search	Act.	Event	Temp.	Avg
VideoLLaMA3-7B	-	-	-	-	49.5	**61.0**	83.0	53.0	61.6	10.7	15.0	9.8	8.2	10.9
ForeSea	✗	✗	✗	3	58.5	58.0	87.0	55.0	64.6	14.8	12.8	12.5	9.8	12.5
	✓	✗	✗	3	53.0	54.0	82.0	56.0	61.3	14.0	11.3	14.2	10.4	12.5
	✗	✓	✗	3	60.0	56.0	85.0	56.0	64.3	15.0	12.5	13.8	11.5	13.2
	✗	✓	✓	3	**61.0**	59.0	85.0	53.0	64.5	15.0	10.1	13.9	9.2	12.1
	✗	✗	✓	3	60.5	60.0	85.0	56.0	65.4	**17.6**	14.1	15.4	8.2	13.8
	✗	✗	✓	5	59.5	**61.0**	**88.0**	**57.0**	**66.4**	15.8	10.6	13.2	8.9	12.1
ForeSea-Global	-	-	-	64*	57.5	58.0	**88.0**	**57.0**	65.1	14.1	**18.4**	19.3	12.7	16.1
ForeSea-Hybrid	✗	✗	✓	2/24*	59.5	59.0	87.0	53.0	64.6	16.9	15.1	**20.1**	**15.0**	**16.8**

localize the evidence and produces wrong answers. In the *search* example where a nonexistent moment is queried, ForeSea correctly identifies the absence of evidence, while VideoLLaMA3 hallucinates a false temporal interval. In the *activity* example, both models fail to localize the moment of interest, but ForeSea still answers correctly by leveraging the retrieved clips. The hardest of all is the *counting* task, where both models under-count the occurrences and fail to follow the output format by providing a list of time intervals, suggesting that counting-based video QA remains a challenging open problem that requires more sophisticated retrieval and reasoning strategies.

5.3 Ablation Studies of ForeSea Configurations

We ablate the key design choices of ForeSea on ForeSeaQA^{MM} in Table 2, including VideoLLaMA3-7B as the no-retrieval baseline. After retrieval, each track is passed to the Video LMM together with optional spatial grounding signals: **Crop** crops the video frames to the tracked bounding box; **Overlay** draws the bounding box on the original (uncropped) frames; **Coords** appends the bounding box coordinates as text in the prompt. Top K controls how many retrieved tracks are concatenated as Video LMM input.

Person-centric retrieval alone outperforms direct video processing. Even without any spatial grounding (no Crop, no Overlay, no Coords), ForeSea with K=3 already surpasses VideoLLaMA3-7B on both accuracy (64.6% vs. 61.6%) and temporal IoU (12.5% vs. 10.9%). This confirms that focusing the Video LMM on a small set of person-centric clips, rather than the full video, is itself a strong inductive bias for forensic search, even before any explicit spatial information is provided.

Text-based coordinate injection for effective spatial grounding. Cropping the video to the bounding box (✓ Crop) actually *hurts* accuracy (61.3%), as it removes the surrounding scene context that the Video LMM relies on for activity and event understanding. Adding a visual bounding box overlay (✓ Overlay) recovers accuracy (64.3%) and improves IoU (13.2%), but the gains are modest. In contrast, passing the bounding box coordinates as text (✓ Coords) achieves the best accuracy–IoU balance (65.4%, 13.8%), and combining Overlay with

Table 3: Comparison across frameworks, retrieval settings, multimodal embeddings, and VLMs. * For comparable conditions with VideoRAG, which uses 64 frames as input to LLaVA-Video, we adopt this setting. MC: multi-choice; TL: temporal localization.

Framework	Setting	VLM	Multimodal Encoder	Multimodal (MM)				Text				Overall	
				MC(accuracy %)		TL(IoU)		MC(accuracy %)		TL(IoU)		MC	TL
				search	avg	search	avg	search	avg	search	avg		
ForeSea	Person Top3	VideoLLaMA3-7B	GCL	60.5	65.4	17.6	13.8	72.0	66.7	28.5	13.3	66.0	13.6
ForeSea-Global	Global Top64	VideoLLaMA3-7B	GCL	57.5	65.1	14.1	16.1	67.5	68.1	25.7	18.6	66.6	17.4
ForeSea-Hybrid	Person Top2 + Global Top24	VideoLLaMA3-7B	GCL	59.5	64.6	16.9	16.8	71.0	69.1	30.5	20.1	66.9	18.4
Different multimodal encoder experiments													
ForeSea	SigLIP	VideoLLaMA3-7B	ViT-SO400M-14-SigLIP-384	59.0	64.5	18.4	13.2	69.0	66.6	28.9	15.8	65.5	14.5
ForeSea	SigLIP2	VideoLLaMA3-7B	ViT-SO400M-16-SigLIP2-512	59.5	64.4	17.4	14.2	70.0	67.3	29.6	15.5	65.8	14.8
ForeSea	EVA-CLIP	VideoLLaMA3-7B	EVA-CLIP EVA02-E-14-plus	59.0	63.3	16.2	12.5	70.0	66.3	28.9	16.4	64.8	14.4
Different VLM experiments													
VideoRAG	CLIP + APE	LLaVA-Video-7B-Qwen2	CLIP-ViT-large-patch14-336	56.5	61.9	3.1	2.8	55.5	63.8	2.2	4.3	62.9	3.5
VideoRAG	CLIP + APE	VideoLLaMA3-7B	CLIP-ViT-large-patch14-336	73.5	67.9	3.3	9.5	72.0	68.4	11.8	10.3	68.1	9.9
ForeSea-Hybrid	Person Top1 + Global Top12*	LLaVA-Video-7B-Qwen2	GCL	77.0	63.3	42.0	13.5	74.0	68.6	30.8	9.3	65.9	11.4
Additional baseline constructed with captions													
Predicted caption	Dense captions by VideoLLaMA3	VideoLLaMA3-7B	Caption only	58.5	52.4	1.1	8.9	52.0	52.2	0.4	10.2	52.3	9.6
Oracle caption	Ground-truth UCA dense captions	VideoLLaMA3-7B	Caption only	71.5	79.9	42.0	28.6	92.0	86.5	28.6	33.7	83.2	31.1

Coords does not improve further (64.5%, 12.1%). This suggests that the Video LMM benefits more from explicit, language-aligned spatial grounding than from visual modifications to the input frames.

More retrieved tracks harm temporal precision. Increasing K from 3 to 5 marginally improves average accuracy (65.4%→66.4%) but consistently degrades temporal IoU (13.8%→12.1%). We therefore adopt K=3 as the best accuracy–localization tradeoff.

Sub-task difficulties. Across all configurations, *Event* accuracy is consistently the highest (82–88%), reflecting that event-level questions can often be answered from a single retrieved clip. *Search* accuracy benefits most from retrieval: `ForeSea` improves from 49.5% (VideoLLaMA3) to 60.5%, confirming that person-centric indexing is the key driver for identity-based queries. *Activity* is the one category where VideoLLaMA3 remains competitive (61.0% vs. 60.0%), likely because activity recognition benefits from broader temporal context that retrieval may truncate. Temporal IoU is uniformly low across all settings (8–12%), indicating that precise temporal grounding remains an open challenge even with person-centric retrieval.

Analysis of `ForeSea` variants. `ForeSea`-Global retrieves full video frames instead of person-centric crops to capture more contextual information. This improves overall performance but reduces *search* accuracy in multiple choice (MC) and temporal localization (TL) IoU (57.5% vs. 60.5% and 14.1% vs. 17.6%), where identity cues are important. This suggests that global indexing benefits scene-level grounding, whereas person retrieval better supports identity-driven queries. By combining both person and global retrieval, `ForeSea`-Hybrid recovers Search performance (59.5% MC and 16.9% TL IoU) while leveraging complementary person- and scene-level context.

5.4 Ablation Studies across Encoders and VLMs

To assess the inherent difficulty and justification of our benchmark (if the tasks can be solved without direct visual grounding), we show text-only baselines

results. Also, To evaluate the generalization of our framework, we conduct extensive experiments varying both the multimodal encoders and the foundational Vision-Language Models (VLMs).

Caption-only baselines. We evaluate two caption-only baselines by replacing ForeSea's video inputs with timestamped captions. The predicted-caption baseline uses dense captions generated by VideoLLaMA3 from uniformly sampled frames, while the oracle-caption baseline uses the ground-truth UCA captions employed to construct ForeSeaQA as an upper bound baseline. As shown in Table 3, predicted captions yield the lowest MC accuracy, although their TL performance remains relatively competitive. This suggests that generated captions often omit fine-grained visual details, while timestamp information provides useful temporal cues. Oracle captions substantially improve overall MC accuracy to 83.2%, confirming the value of high-quality semantic descriptions. However, their TL performance remains limited, indicating that precise temporal grounding is challenging even with oracle captions.

Sensitivity to the multimodal encoder. We next investigate whether ForeSea is sensitive to the choice of multimodal retrieval encoder. In addition to our default GCL encoder, we evaluate SigLIP, SigLIP2, and EVA-CLIP under the same ForeSea retrieval and VideoLLaMA3 inference pipeline. The four encoders produce similar overall performance. This show that that the effectiveness of ForeSea does not depend on a particular multimodal embedding model.

Effect of the VLM backbone. To disentangle the contribution of the RAG frame work and downstream VLM, we cross-evaluate VideoRAG [25] and ForeSea using both VideoLLaMA3 and LLaVA-Video. With the same VideoLLaMA3 backbone, ForeSea-Hybrid achieves an overall TL IoU of 18.4%, compared with 9.9% for VideoRAG, while maintaining comparable MC accuracy (66.9% versus 68.1%). Thus, ForeSea provides nearly a twofold improvement in temporal localization without relying on a stronger VLM. The improvement is also preserved when using LLaVA-Video. Under this backbone, ForeSea-Hybrid improves the overall MC accuracy from 62.9% to 65.9% and the TL IoU from 3.5% to 11.4% compared with VideoRAG. Although LLaVA-Video is substantially weaker than VideoLLaMA3 on temporal localization, ForeSea consistently provides a large relative improvement. These results show that the temporal localization gains primarily arise from the proposed framework rather than from a specific VLM backbone.

5.5 Efficiency Analysis

As shown in Table 4, `ForeSea` achieves lower total latency than all baselines while maintaining higher accuracy. By retrieving only the most relevant person-centric clips, `ForeSea` reduces the number of frames fed to the Video LMM, directly lowering TTFT and generation time compared to VideoLLaMA3 (which processes the full video). `ForeSea` completes inference in 2.6 s total (1.7 s TTFT) while achieving the best $\texttt{ForeSeaQA}^{\texttt{MM}}$ accuracy (65.4%). In contrast, T^* incurs

Table 4: Inference latency on ForeSeaQA. TTFT stands for time to first token. Retrieval, generation, and total time are in seconds; accuracy and IoU in %.

Method	Latency (s)			ForeSeaQA^{MM}	
	Retrieval	Generation$_{(TTFT)}$	Total	Acc	IoU
Qwen2.5-VL-7B-Instruct [3]	**0.0**	2.1$_{(1.7)}$	2.1	58.9	8.1
VideoLLaMA3-7B [47]	**0.0**	3.8$_{(3.6)}$	3.8	61.6	10.9
VideoRAG [25] LLaVA-Video-7B-Qwen2	2.4	2.8$_{(2.3)}$	5.2	61.9	2.8
T* [41] Qwen2.5-VL-7B-Instruct	6.8	**0.9**$_{(0.6)}$	7.6	41.1	4.9
ForeSea	0.5	2.1$_{(1.7)}$	2.6	**65.4**	13.8
ForeSea-Global	0.5	**0.9**$_{(0.6)}$	**1.4**	65.1	**16.1**

Table 5: Performance on open-domain long video benchmarks. All numbers are reported from the original papers.

Model	Param	Year	VideoMME	MLVU
LongVU [32]	7B	2024	–	65.4
LLaVA-Video [19]	7B	2024	56.6	64.7
TimeMarker [8]	7B	2024	57.3	49.2
InternVL2.5 [9]	7B	2024	56.3	64.0
Qwen2.5VL [3]	7B	2025	65.1	70.2
VideoLLaMA3 [47]	7B	2025	66.2	73.0
LLaVA-Video + Video-RAG [25]	7B	2024	58.7	72.4
SALOVA-7B [17]	7B	2025	53.1	–
MemVid-7B [43]	7B	2025	63.7	58.1
GPT-4o + T* [41]	>7B	2025	56.5	–
LLaVA-OneVision-72B + T* [41]	72B	2025	59.0	–
ForeSea (Ours)	7B	–	65.6	73.0

the highest retrieval latency (6.8 s) despite fast generation, and VideoRAG adds overhead from its dedicated retrieval pipeline (2.4 s retrieval).

5.6 Comparison on Existing Benchmarks

To assess generalization beyond the surveillance domain, we evaluate ForeSea on three widely used long video benchmarks: VideoMME [12] MLVU [54], and LongVideoBench [38]. [8] For these benchmarks, ForeSea adapts its database construction: instead of person-centric clips, frames are sampled uniformly at 1 FPS and indexed at the frame level. The backbone Video LMM, VideoLLaMA3, supports up to 180 input frames; ForeSea uses at most 90 frames per query (top-60 retrieved + 30 uniformly sampled from the full video). Despite using only half as many frames, ForeSea achieves comparable performance across all three benchmarks and substantially outperforms prior Video-RAG approaches, as shown in Table 5.

6 Conclusion

We introduced ForeSea, a novel Video-RAG framework for forensic search in human surveillance video. ForeSea is, to our knowledge, the first system to handle complex multimodal (image+text) queries and return timestamped, evidence-linked answers, overcoming the limitations of text-only retrieval. To validate this, we also developed ForeSeaQA, the first benchmark for evaluating such temporally-grounded multimodal queries. Our experiments demonstrate that ForeSea's pipeline achieves significant gains in both QA accuracy and temporal IoU over strong baselines. Furthermore, we show our framework's extensibility beyond surveillance, demonstrating its effectiveness on general video understanding tasks. This work provides a robust framework and a critical evaluation tool, marking a significant step forward in practical AI forensic analysis.

[8] LongVideoBench results are provided in the supplementary material.

References

1. Ataallah, K., Gou, C., Abdelrahman, E., Pahwa, K., Ding, J., Elhoseiny, M.: Infinibench: A comprehensive benchmark for large multimodal models in very long video understanding. EMNLP (2025)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
4. Bergmann, P., Meinhardt, T., Leal-Taixé, L.: Tracking without bells and whistles. In: ICCV (2019)
5. Cai, Q., Liang, H., Dong, H., Qiang, M., An, R., Han, Z., Zhu, Z., Cui, B., Zhang, W.: Lovr: A benchmark for long video retrieval in multimodal contexts. arXiv preprint arXiv:2505.13928 (2025)
6. Cao, M., Bai, Y., Zeng, Z., Ye, M., Zhang, M.: An empirical study of clip for text-based person search. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 465–473 (2024)
7. Chen, G., Zheng, Y.D., Wang, J., Xu, J., Huang, Y., Pan, J., Wang, Y., Wang, Y., Qiao, Y., Lu, T., et al.: Videollm: Modeling video sequence with large language models. arXiv preprint arXiv:2305.13292 (2023)
8. Chen, S., Lan, X., Yuan, Y., Jie, Z., Ma, L.: Timemarker: A versatile video-llm for long and short video understanding with superior temporal localization ability (2024), https://arxiv.org/abs/2411.18211
9. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
10. Cheng, J., Ge, Y., Wang, T., Ge, Y., Liao, J., Shan, Y.: Video-holmes: Can mllm think like holmes for complex video reasoning? arXiv preprint arXiv:2505.21374 (2025)
11. Feng, K., Gong, K., Li, B., Guo, Z., Wang, Y., Peng, T., Wu, J., Zhang, X., Wang, B., Yue, X.: Video-r1: Reinforcing video reasoning in mllms. arXiv preprint arXiv:2503.21776 (2025)
12. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., Chen, P., Li, Y., Lin, S., Zhao, S., Li, K., Xu, T., Zheng, X., Chen, E., Shan, C., He, R., Sun, X.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis (2025), https://arxiv.org/abs/2405.21075
13. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: Proceedings of the IEEE international conference on computer vision. pp. 5267–5275 (2017)

14. Hong, W., Yu, W., Gu, X., Wang, G., Gan, G., Tang, H., Cheng, J., Qi, J., Ji, J., Pan, L., et al.: Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. arXiv e-prints pp. arXiv–2507 (2025)
15. Jeong, S., Kim, K., Baek, J., Hwang, S.J.: Videorag: Retrieval-augmented generation over video corpus. ACL (2025)
16. Jocher, G.: Ultralytics yolov5 (2020). https://doi.org/10.5281/zenodo.3908559, https://github.com/ultralytics/yolov5
17. Kim, J., Kim, H., Lee, H., Ro, Y.M.: Salova: Segment-augmented long video assistant for targeted retrieval and routing in long-form video analysis. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3352–3362 (2025)
18. Lee, J., Cho, J., Park, H., Hayat, M., Hwang, K., Porikli, F., Choi, S.: Generalized contrastive learning for universal multimodal retrieval (2025)
19. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer (2024), https://arxiv.org/abs/2408.03326
20. Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., Ma, Z., Li, C.: Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. arXiv preprint arXiv:2407.07895 (2024)
21. Li, S., Xiao, T., Li, H., Zhou, B., Yue, D., Wang, X.: Person search with natural language description. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1970–1979 (2017)
22. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. arXiv preprint arXiv:2311.10122 (2023)
23. Liu, B., Qiao, P., Ma, M., Zhang, X., Tang, Y., Xu, P., Liu, K., Yuan, T.: Surveillancevqa-589k: A benchmark for comprehensive surveillance video-language understanding with large models. arXiv preprint arXiv:2505.12589 (2025)
24. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: Clip4clip: An empirical study of clip for end to end video clip retrieval. arXiv preprint arXiv:2104.08860 (2021)
25. Luo, Y., Zheng, X., Yang, X., Li, G., Lin, H., Huang, J., Ji, J., Chao, F., Luo, J., Ji, R.: Video-rag: Visually-aligned retrieval-augmented long video comprehension. NeurIPS (2025)
26. Mao, M., Perez-Cabarcas, M.M., Kallakuri, U., Waytowich, N.R., Lin, X., Mohsenin, T.: Multi-rag: A multimodal retrieval-augmented generation system for adaptive video understanding. arXiv preprint arXiv:2505.23990 (2025)
27. Pang, J., Qiu, L., Li, H., et al.: Quasi-dense similarity learning for multiple object tracking. In: CVPR (2021)
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)
29. Ren, X., Xu, L., Xia, L., Wang, S., Yin, D., Huang, C.: Videorag: Retrieval-augmented generation with extreme long-context videos. arXiv preprint arXiv:2502.01549 (2025)
30. Sagare, S.R., Ullegaddi, P., Sarabhai, K., SA, R.K., et al.: Videorag: Scaling the context size and relevance for video question-answering. In: INLG (2024)
31. Shao, Y., He, H., Li, S., Chen, S., Long, X., Zeng, F., Fan, Y., Zhang, M., Yan, Z., Ma, A., et al.: Eventvad: Training-free event-aware video anomaly detection. arXiv preprint arXiv:2504.13092 (2025)

32. Shen, X., Xiong, Y., Zhao, C., Wu, L., Chen, J., Zhu, C., Liu, Z., Xiao, F., Varadarajan, B., Bordes, F., Liu, Z., Xu, H., Kim, H.J., Soran, B., Krishnamoorthi, R., Elhoseiny, M., Chandra, V.: Longvu: Spatiotemporal adaptive compression for long video-language understanding (2024), https://arxiv.org/abs/2410.17434
33. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6479–6488 (2018)
34. Wang, H., Xu, Z., Cheng, Y., Diao, S., Zhou, Y., Cao, Y., Wang, Q., Ge, W., Huang, L.: Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models. arXiv preprint arXiv:2410.03290 (2024)
35. Wang, W., He, Z., Hong, W., Cheng, Y., Zhang, X., Qi, J., Gu, X., Huang, S., Xu, B., Dong, Y., Ding, M., Tang, J.: Lvbench: An extreme long video understanding benchmark (2025), https://arxiv.org/abs/2406.08035
36. Wang, X., Zhang, Y., Zohar, O., Yeung-Levy, S.: Videoagent: Long-form video understanding with large language model as agent. In: European Conference on Computer Vision. pp. 58–76. Springer (2024)
37. Wang, Y., Wang, Z., Xu, B., Du, Y., Lin, K., Xiao, Z., Yue, Z., Ju, J., Zhang, L., Yang, D., et al.: Time-r1: Post-training large vision language model for temporal video grounding. arXiv preprint arXiv:2503.13377 (2025)
38. Wu, H., Li, D., Chen, B., Li, J.: Longvideobench: A benchmark for long-context interleaved video-language understanding (2024), https://arxiv.org/abs/2407.15754
39. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z.: Qwen3 technical report (2025), https://arxiv.org/abs/2505.09388
40. Yang, Z., Wang, S., Zhang, K., Wu, K., Leng, S., Zhang, Y., Li, B., Qin, C., Lu, S., Li, X., et al.: Longvt: Incentivizing" thinking with long videos" via native tool calling. arXiv preprint arXiv:2511.20785 (2025)
41. Ye, J., Wang, Z., Sun, H., Chandrasegaran, K., Durante, Z., Eyzaguirre, C., Bisk, Y., Niebles, J.C., Adeli, E., Fei-Fei, L., et al.: Re-thinking temporal search for long-form video understanding. In: CVPR (2025)
42. Ye, M., Liu, W., He, P.: Vera: Explainable video anomaly detection via verbalized learning of vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2025)
43. Yuan, H., Liu, Z., Qin, M., Qian, H., Shu, Y., Dou, Z., Wen, J.R., Sebe, N.: Memory-enhanced retrieval augmentation for long video understanding. arXiv preprint arXiv:2503.09149 (2025)
44. Yuan, H., Ni, J., Liu, Z., Wang, Y., Zhou, J., Liang, Z., Zhao, B., Cao, Z., Dou, Z., Wen, J.R.: Momentseeker: A task-oriented benchmark for long-video moment retrieval. arXiv preprint arXiv:2502.12558 (2025)
45. Yuan, T., Zhang, X., Liu, K., Liu, B., Chen, C., Jin, J., Jiao, Z.: Towards surveillance video-and-language understanding: New dataset, baselines, and challenges (2023), https://arxiv.org/abs/2309.13925
46. Zanella, L., Menapace, W., Mancini, M., Wang, Y., Ricci, E.: Harnessing large language models for training-free video anomaly detection. In: CVPR (2024)

47. Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., et al.: Videollama 3: Frontier multimodal foundation models for image and video understanding. arXiv preprint arXiv:2501.13106 (2025)
48. Zhang, G., Fok, M.L.A., Ma, J., Xia, Y., Cremers, D., Torr, P., Tresp, V., Gu, J.: Localizing events in videos with multimodal queries (2024), https://arxiv.org/abs/2406.10079
49. Zhang, H., Xu, X., Wang, X., Zuo, J., Huang, X., Gao, C., Zhang, S., Yu, L., Sang, N.: Holmes-vau: Towards long-term video anomaly understanding at any granularity. In: CVPR (2025)
50. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. In: European conference on computer vision. pp. 1–21. Springer (2022)
51. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Llava-video: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024)
52. Zhao, X., Zhang, C., Guo, P., Li, W., Chen, L., Zhao, C., Huang, S.: Smarthome-bench: A comprehensive benchmark for video anomaly detection in smart homes using multi-modal large language models. In: CVPR (2025)
53. Zhou, J., Liu, Z., Xiao, S., Zhao, B., Xiong, Y.: VISTA: Visualized text embedding for universal multi-modal retrieval. In: ACL (Aug 2024)
54. Zhou, J., Shu, Y., Zhao, B., Wu, B., Liang, Z., Xiao, S., Qin, M., Yang, X., Xiong, Y., Zhang, B., Huang, T., Liu, Z.: Mlvu: Benchmarking multi-task long video understanding (2025), https://arxiv.org/abs/2406.04264
55. Zhou, X., Larintzakis, K., Guo, H., Zimmer, W., Liu, M., Cao, H., Zhang, J., Lakshminarasimhan, V., Strand, L., Knoll, A.C.: Tumtraffic-videoqa: A benchmark for unified spatio-temporal video understanding in traffic scenes. ICML (2025)
56. Zhou, X., Koltun, V., Krähenbühl, P.: Tracking objects as points. In: ECCV (2020)
57. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)