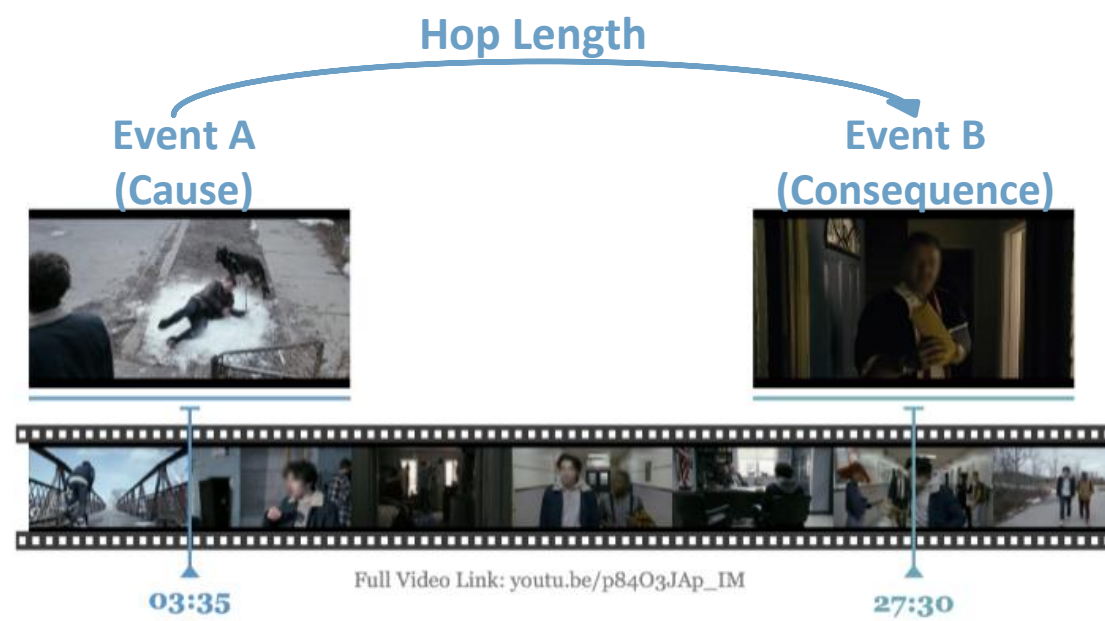


Multi-Hop Reasoning in Long Videos



Videos can be very long (multiple hours):

- How to connect **temporally distant events** in long videos?

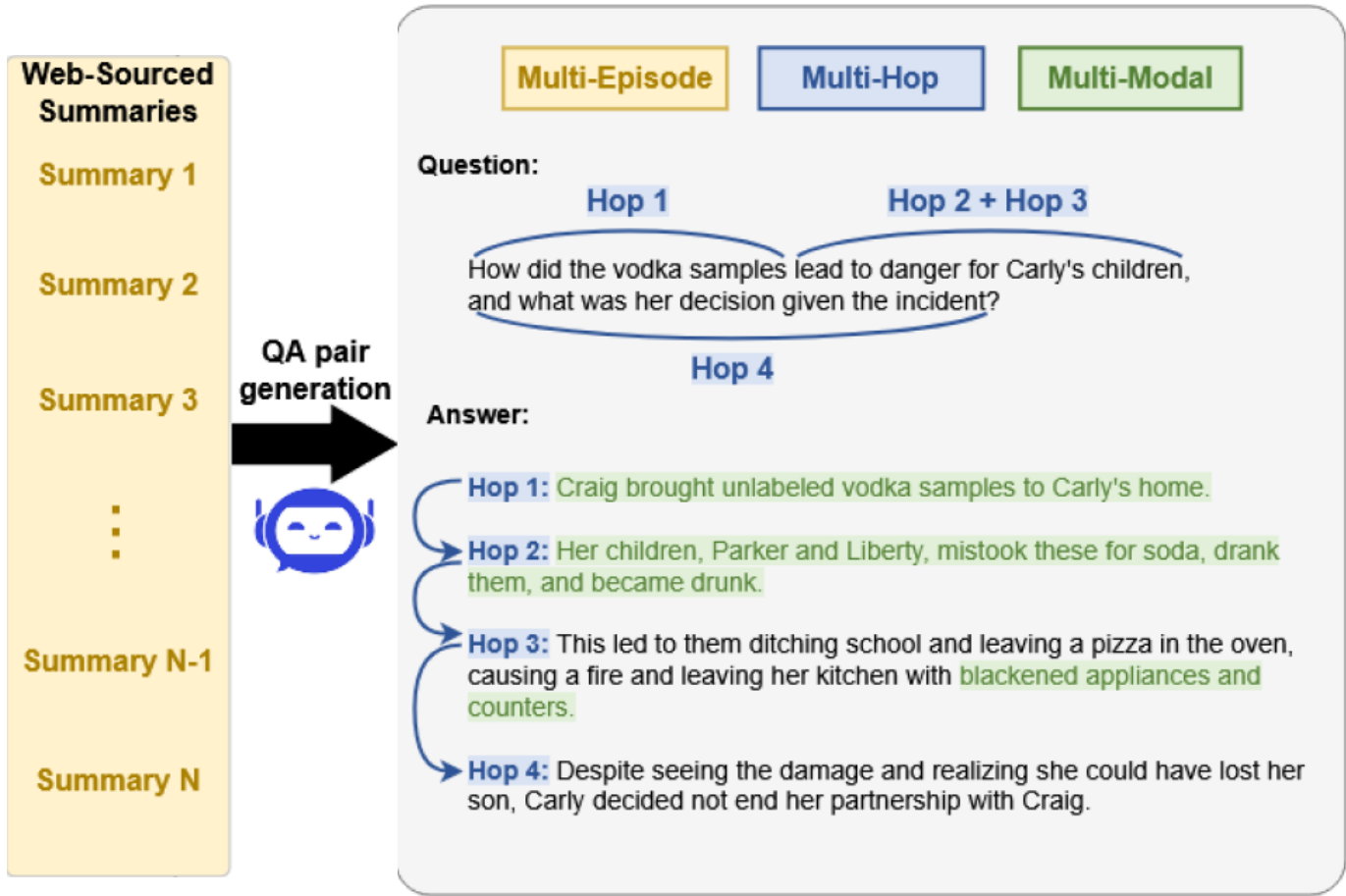
=> **Multi-Hop Question Answering Benchmarks**

Data Generation Pipeline of SagaQA

SagaQA: 200 questions spanning 4 years of the TV Show *As the World Turns*

Each **Question-Answer (QA) pair** targets a sequence of **20 consecutive episodes**

We generate **QA pairs** given human annotated summaries of the 20 episode,



Multi-Episode: The questions cannot be answered using a single episode

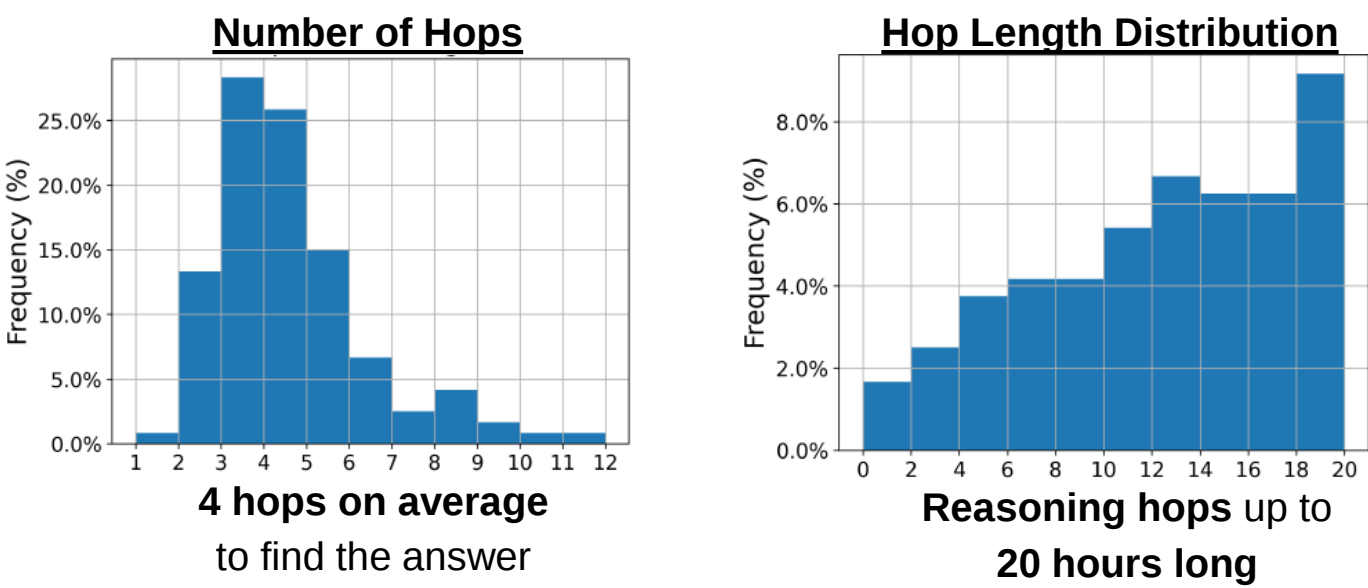
Multi-Hop: The questions must integrate information from multiple audio-visual segments through a series of dependent reasoning steps that progressively build toward the final answer.

Multi-Modal: The questions cannot be answered from dialogue alone and require analyzing visual elements from the video.

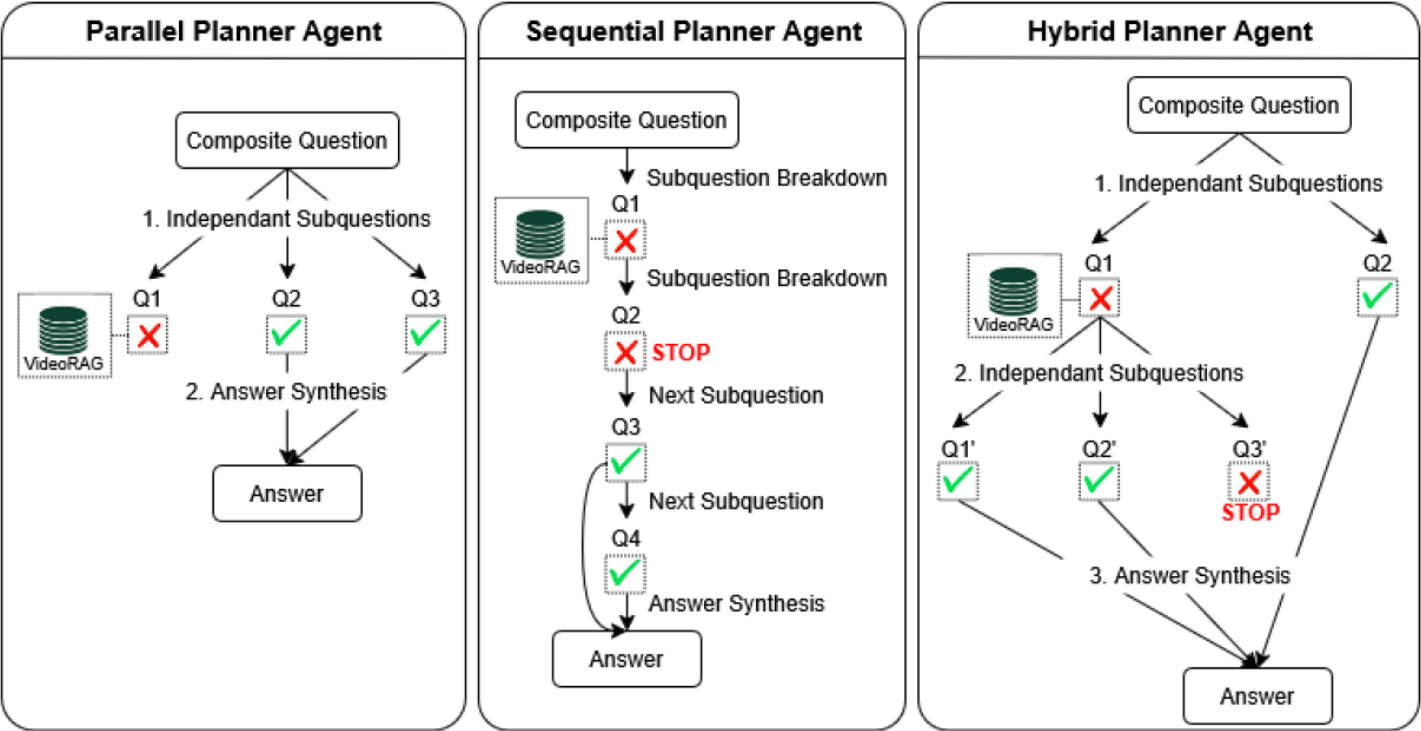
Statistics Distribution of SagaQA

Dataset	Hop Length	Reasoning
Short Hops		
CausalStep (Li et al., 2025)	sec-min	Narrative
Video-Holmes (Cheng et al., 2025)	sec-min	Narrative
VRBench (Yu et al., 2025a)	Up to 40min	Narrative
Long Hops		
Video-BrowseComp (Liang et al., 2026)	Search-Based	Fact-Finding
Ego-LifeQA (Yang et al., 2025a)	Up to 24h+	Temporal
SagaQA (ours)	Up to 20h	Narrative

The **hops in SagaQA** are **longer (up to 20 hours)** and **more challenging** than other datasets by targeting the whole **narrative**



Video Agentic Planning



Parallel Planner: Solves a question by answering independent subquestions in parallel. (+ **Exploration**, – **Exploitation**)

Sequential Planner: Solves a question through a sequence of dependent reasoning steps, one subquestion at a time. (– **Exploration**, + **Exploitation**)

Hybrid Planner: Combines sequential planning with parallel subquestion execution. (+ **Exploration**, + **Exploitation**)

Results

	Answer Evaluation				Episode Grounding		
	r1	r2	rlsum	METEOR	Precision	Recall	F1
Non-Planning							
TextRAG	36.30	6.51	33.09	23.49	28.13	11.14	14.92
VideoRAG	36.71	6.72	33.27	24.42	36.91	30.46	28.54
Parallel Planners							
EGAgent (Qwen3-30B-Instruct)	33.01	4.25	29.41	20.38	38.89	33.33	32.86
Parallel Planner (Prompt 1)	35.00	5.93	31.93	23.42	50.03	26.35	30.88
Parallel Planner (Prompt 2)	36.06	6.15	32.98	24.10	48.06	30.93	34.26
Sequential Planners							
VideoExplorer (7B)	27.18	10.67	23.89	18.94	33.33	10.42	15.83
Sequential Planner + VideoRAG	36.94	6.67	33.35	24.22	46.76	30.43	32.68
Hybrid Planners							
Hybrid Planner + VideoRAG	35.99	6.32	32.58	24.07	45.47	44.11	41.02

Assess the generated answer with respect to reference

Assess the episodes retrieved by the agent to ground the answer

Hybrid Planner: Best trade-off between **Exploration (Recall)** and **Exploitation (Precision)**

Post-analysis of retrieved video segments remains a bottleneck: ROUGE and METEOR scores are low

Takeaways

SagaQA for **long hop reasoning** in **challenging** TV show narratives

Structured Hybrid Video Planning for better **Exploration vs. Exploitation**

Current methods still fail to **integrate information across long TV shows**