USENIX NSDI 2026

Microsoft

AVA: Towards Agentic Video Analytics with Vision Language Models

Yuxuan Yan[1], Shiqi Jiang[2], Ting Cao[3], Yifan Yang[2], Qianqian Yang[1]
Yuanchao Shu[1] , Yuqing Yang[2], Lili Qiu[2]

[1]Zhejiang University [2]Microsoft Research [3]Tsinghua University

Video Analytics

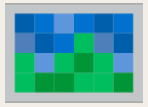
Video Analytics is the process of using algorithms to automatically extract and understand information from video, and generate useful insights.

1

Signal & Motion Processing
1960s – early 1990s

Operating on raw pixel signals

Horn-Schunck Optical Flow - 1981
Lucas-Kanade Tracker - 1981
Canny Edge Detector - 1986
Kalman Filter Tracking - 1980s+
GMM Background Modeling - 1999



2

Engineered Features + ML
mid 1990s — 2011

Operating on feature spaces

Viola-Jones Face Detection – 2001
Bag-of-Visual-Words - 2003
HOG + SVM Pedestrians - 2005
Deformable Part Model - 2008
Dense Trajectories - 2011

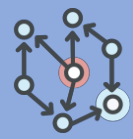


3

AI Era
2012 — Now

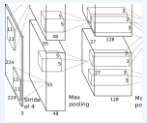
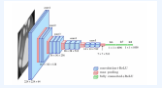
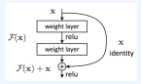
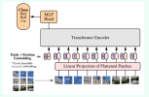
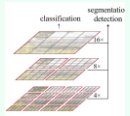
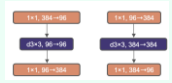
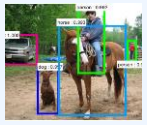
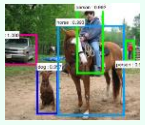
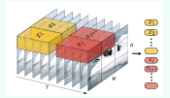
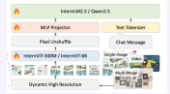
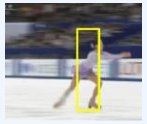
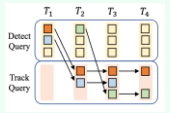
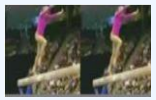
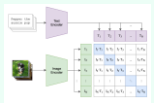
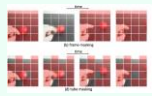
Operating on learned representations

AlexNet – 2012
YOLO / Faster R-CNN – 2016
I3D – 2017
ByteTrack – 2022
SAM – 2023
GPT4-o / Gemini 1.5 -2024

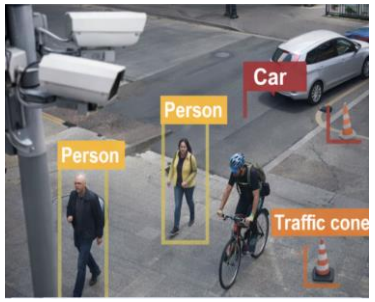

Video Analytics with Artificial Intelligence

Evolution of AI Models for Video Analytics

	2012-2018 CNN			2019-2022 Transformer			2023-now Foundation model		
BACKBONE	 AlexNet 2012	 VGGNet 2015	 ResNet 2015	BACKBONE	 VIT 2020	 Swin-Trans. 2021	 ConvNeXt 2022	FOUNDATION MODEL	 SAM 2023	 SAM2 2024	 DINOv2 2023
DTECTION	 Faster R-CNN 2015	 YOLO-v1 2016	 SSD 2016	VIDEO BACKBONE	 TimeSformer 2021	 Video Swin 2021	 ViViT 2021	VISION LANGUAGE MODEL	 Video-LLaVA 2023	 Qwen2-VL 2024	 ItemVL-2 2024
TRACKING	 SiamFC 2016	 SORT 2016	 DeepSORT 2017	TRACKING	 TransT 2021	 ByteTrack 2022	 MoTR 2022	MULTIMODAL FLAGSHIP	 GPT4-o 2024	 Gemini 1.5 pro 2024	 Doubao 1.5 pro 2025
ACTION RECOGNITION	 2S-CNN 2014	 C3D 2015	 I3D 2017	VISION LANGUAGE	 CLIP 2021	 MAE 2021	 VideoMAE 2022	AGENTIC INTELLIGENCE	 VCA 2024	 AVA 2025	 StreamClaw 2026

Intelligence Level of Video Analytics

Level 1
Basic Perception



- **Definition**
  Extract structured atomic facts from each video frame

- **Input → Output**
  Frame → Objects (class + bbox + ID + timestamp)

- **Examples**
  Scene: Street camera
  - Car
  - Person
  - …

Limitations
Only answers “what is visible” in pixels, not “what is happening”

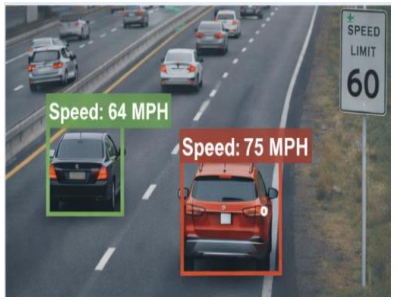
2012-2018
CNN

2019-2022
Transformer

2023-now
Foundation model

Intelligence Level of Video Analytics

Level 2
Event Detection



- **Definition**
  Recognize motion events across frames

- **Input → Output**
  Frame * N → Events (type + time + actors)

- **Examples**
  Scene: Traffic camera
  - Speeding
  - Crowd anomaly
  - …

- **Limitations**
  Events and pipelines are pre-defined and manually configured

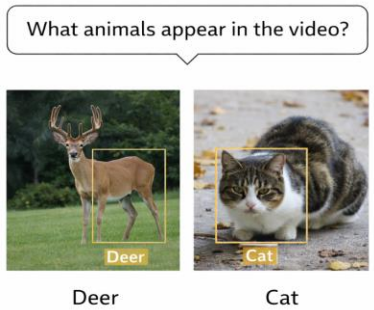
2012-2018
CNN

2019-2022
Transformer

2023-now
Foundation model

Intelligence Level of Video Analytics

Level 3
Language Query



- **Definition**
  Search and answer videos with free-form natural language

- **Input → Output**
  Video + Text query → Text answer / Video clip

- **Examples**
  Scene: Wildlife video
  - Deer 🦌 @02:14 — 02:36
  - Animals 🐱 🐶 🐰
  - …

- **Limitations**
  Generalizes detection but lacks reasoning capability

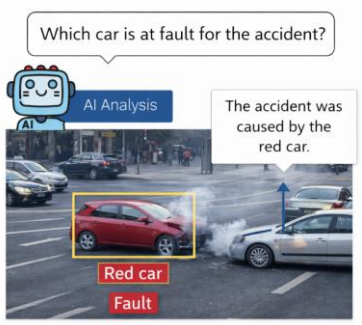
2012-2018
CNN

2019-2022
Transformer

2023-now
Foundation model

Intelligence Level of Video Analytics

Level 4
Open-ended Reasoning



- **Definition**
  Answer any open-ended question through reasoning

- **Input → Output**
  Video + Open question → Multi-step reasoning

- **Examples**
  Scene: Traffic accident
  - "Which car is at fault?" → Red car ran red light
  - ...

Status
Emerging — VLM demonstrate strong visual capabilities

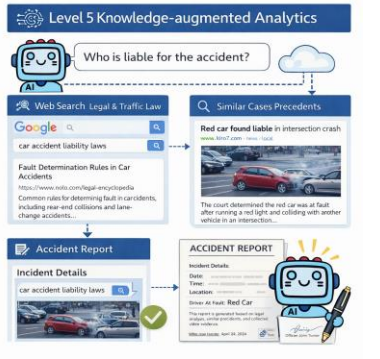
2012-2018
CNN

2019-2022
Transformer

2023-now
Foundation model

Intelligence Level of Video Analytics

Level 5
Knowledge-augmented Analytics



- **Definition**
  Combine video/external knowledge/tools for e2e workflow

- **Input → Output**
  Video + Goal + Agent → Action/Report

- **Examples**
  Issue an accident liability report.
  - Document with cited laws, relevant precedents, fault determination, and settlement recommendation.

Research stage
Early exploration on text-only agent – OpenClaw, Harmes

2012-2018
CNN

2019-2022
Transformer

2023-now
Foundation model

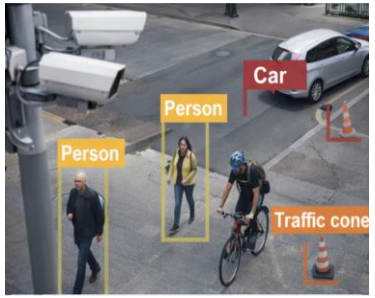
Video Analytics with Different Intelligence Levels

Past Work

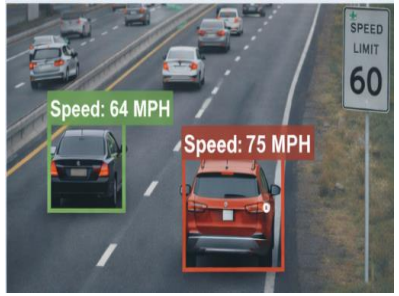
AVA

Future Work

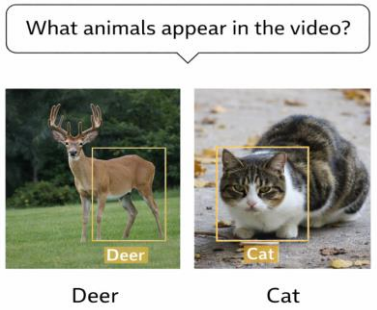
Level 1 Basic Perception



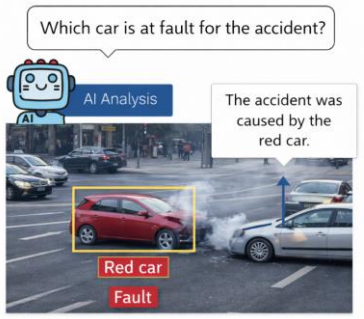
Level 2 Event Detection



Level 3 Language Query



Level 4 Open-ended Reasoning



Level 5 Knowledge-augmented Analytics

2012-2018
CNN

2019-2022
Transformer

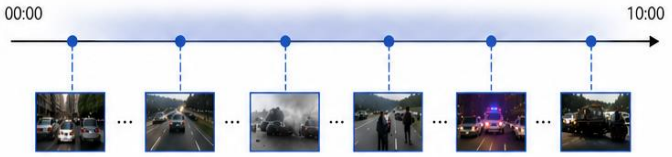
2023-now
Foundation model

Challenge 1 . Struggle to Handle Open-ended Tasks

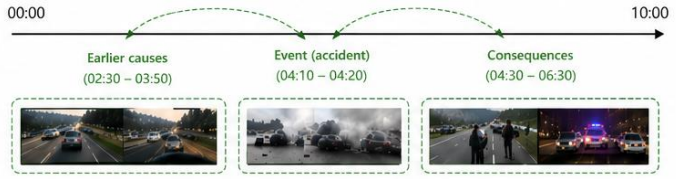
Video evidence distributes differently across question types.

- **Key info-centric**
  - Q *When did the accident happen?*
  - **Direct evidence**
  - *The video evidence is strongly correlated with the query semantics.*

- **Summary**
  - Q *What's the video talking about?*
  - **Distributed info**
  - *Understanding requires information spread across the whole video*



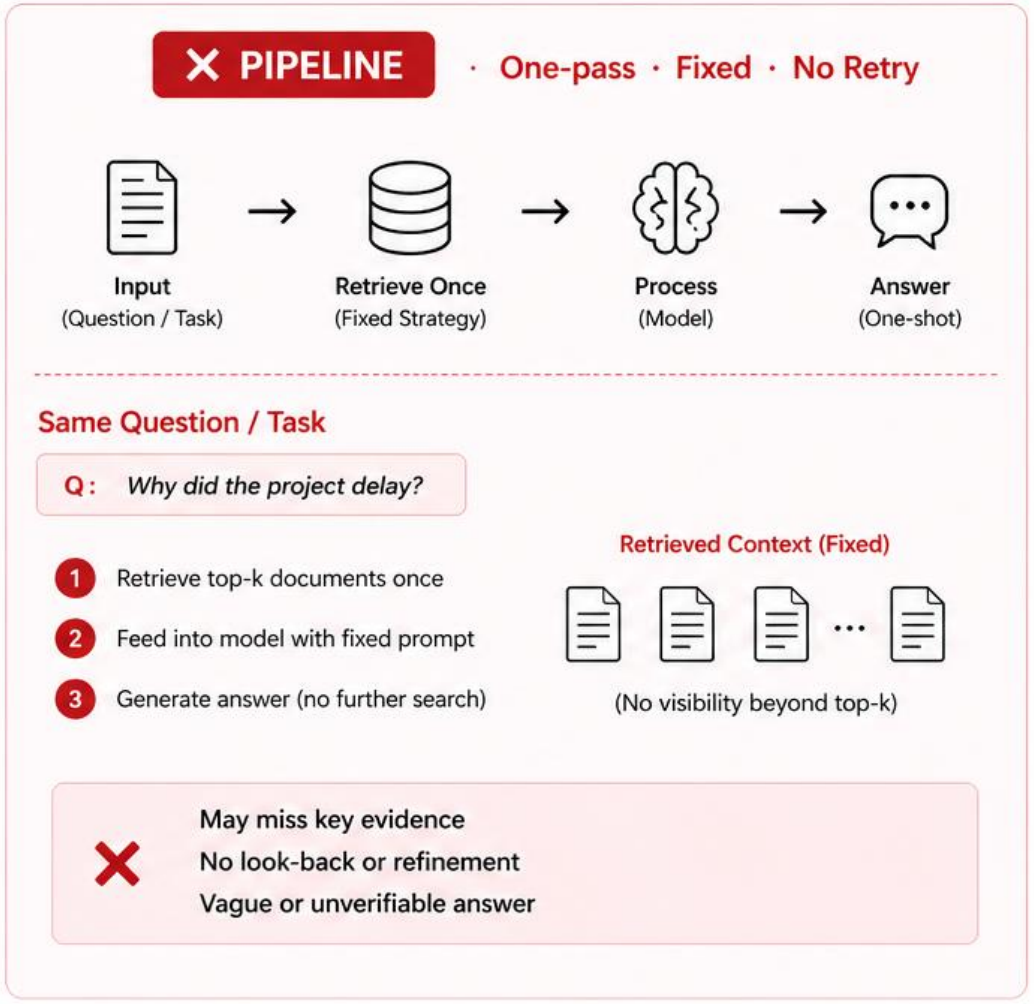
- **Reasoning**
  - Q *When cause the chain of accident?*
  - **Multi-hop**
  - *The answer requires reasoning across multi-steps*



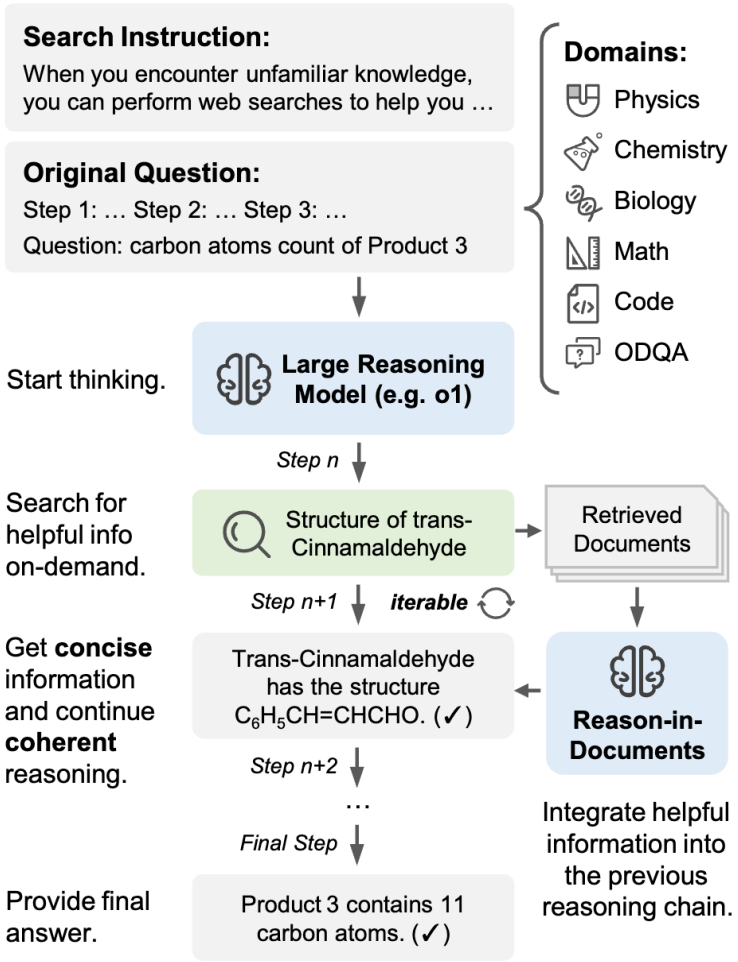
- **Others**
  - **Complex**

Observation 1. Pipelines Run Once. Agents Iterate.

Iteratively plan, retrieve, reason — agents close what pipelines leave open.



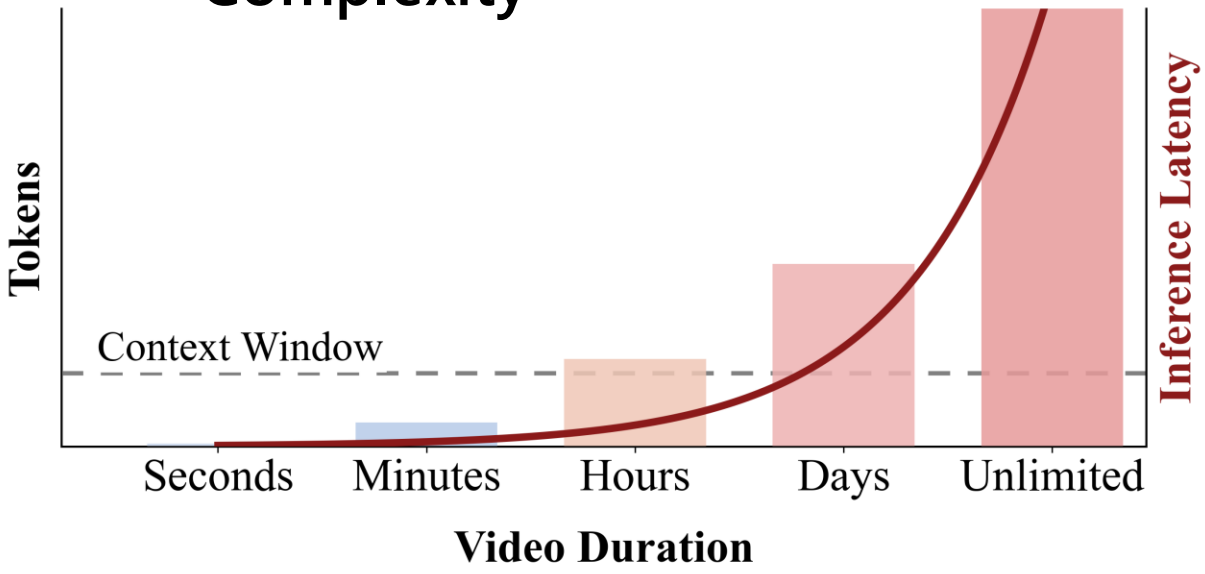
Agentic search on web searching



Challenge 2 . Overhead to Handle Ultra-long Videos

When videos duration grow, tokens and computation complexity explodes

⚠ Tokens ⚠ Complexity



⚠ Limited Context Window

Model	Context Window (tokens)	Release Date
GPT-4o	128K	May 2024
Gemini 1.5 Pro	1M	May 2024
Gemini 1.5 Flash	1M	May 2024
Claude 3.5 Sonnet	200K	Jun 2024
Claude 3 Opus	200K	Mar 2024
Llama 3.2 Vision (90B)	128K	Sep 2024
Qwen2-VL (72B)	128K	Aug 2024
InternVL 2.5 (78B)	128K	Jan 2025
Doubao 1.5 Vision Pro	128K	May 2024
Mistral Large 2	32K	Jul 2024

Example

1 Hour Video —1fps sample→ 3600 Frames —GPT→ ~600k tokens >> 128k

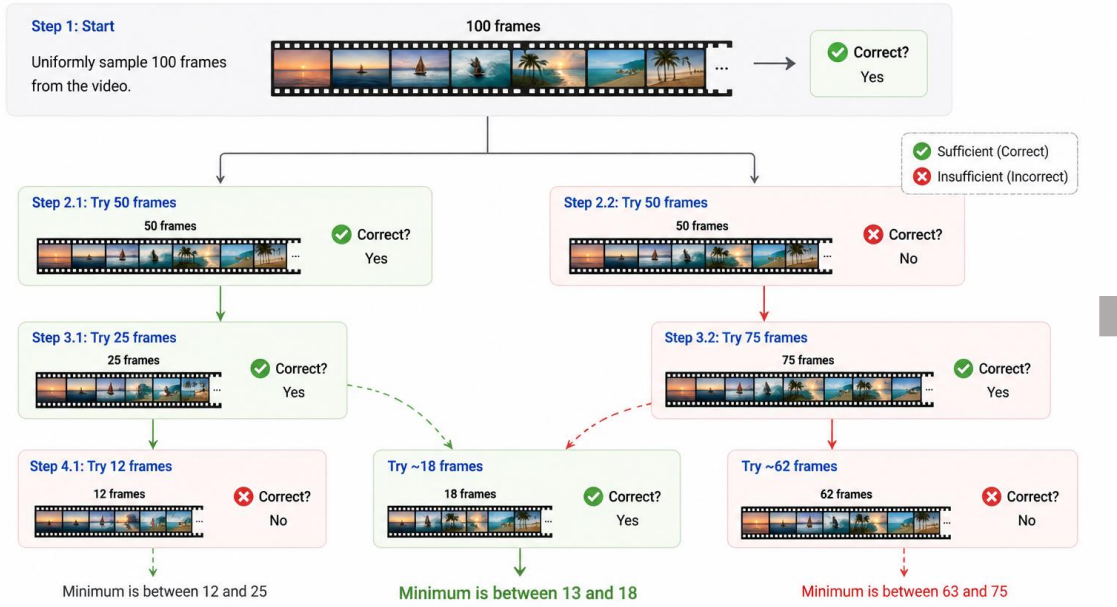
GPT4-o $2.5 / M tokens

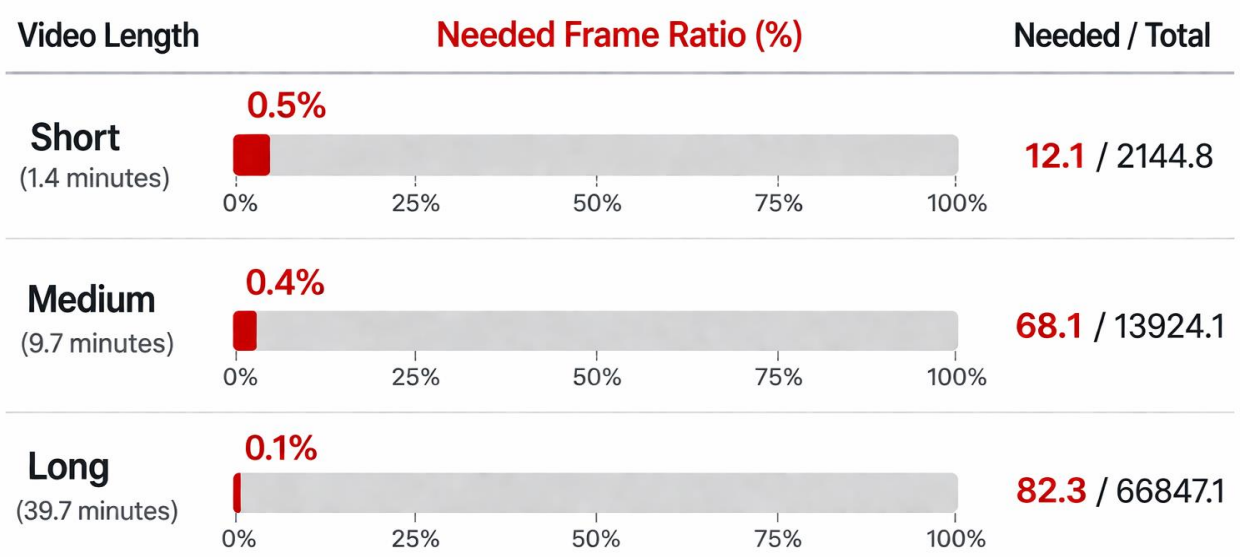
GPT5.5 $5 / M tokens

Observation 2. Useful information is sparse

Despite the token explosion, for most query, < 1% of frames hold answer.

QwenVL-2 on Video-MME



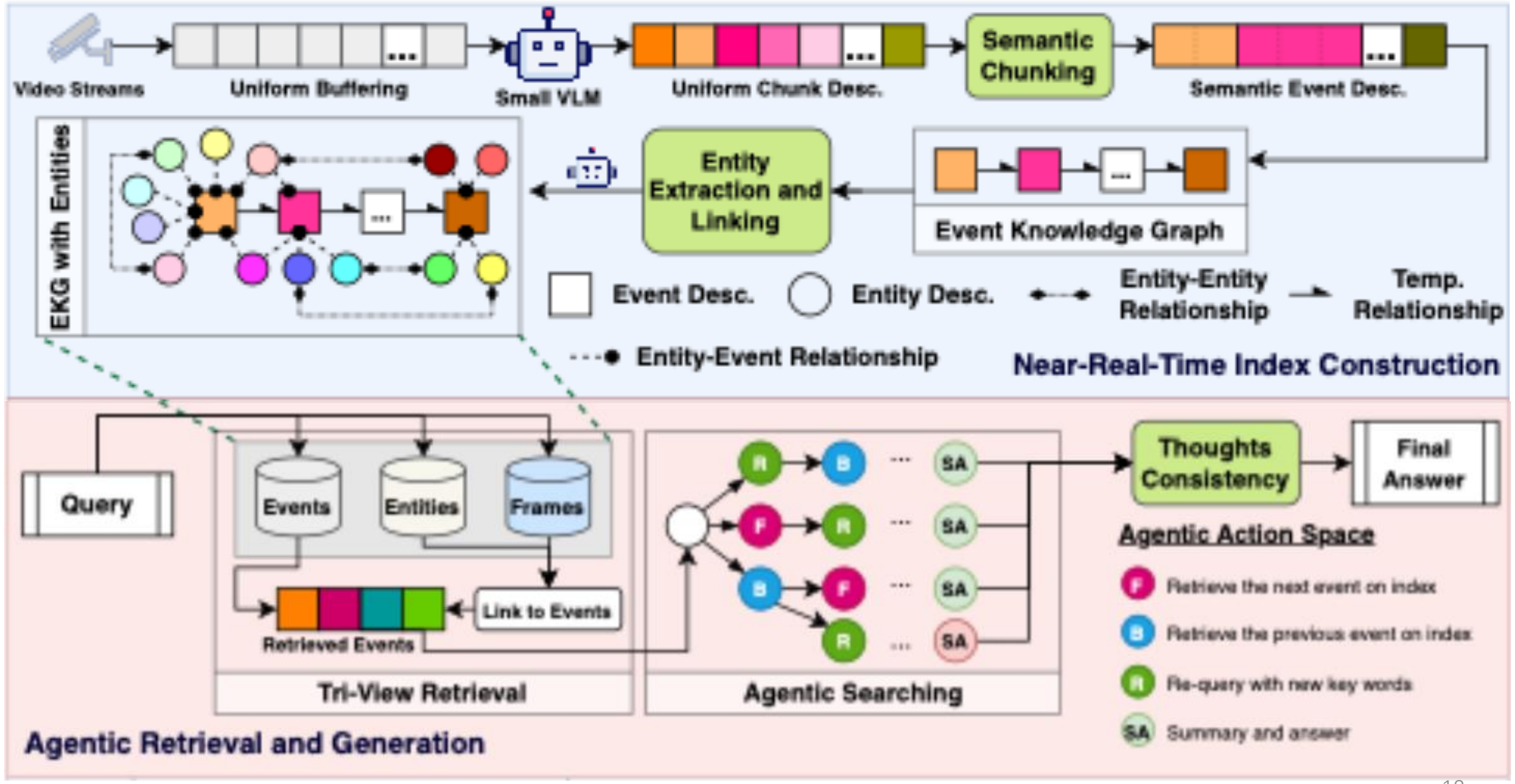


Example

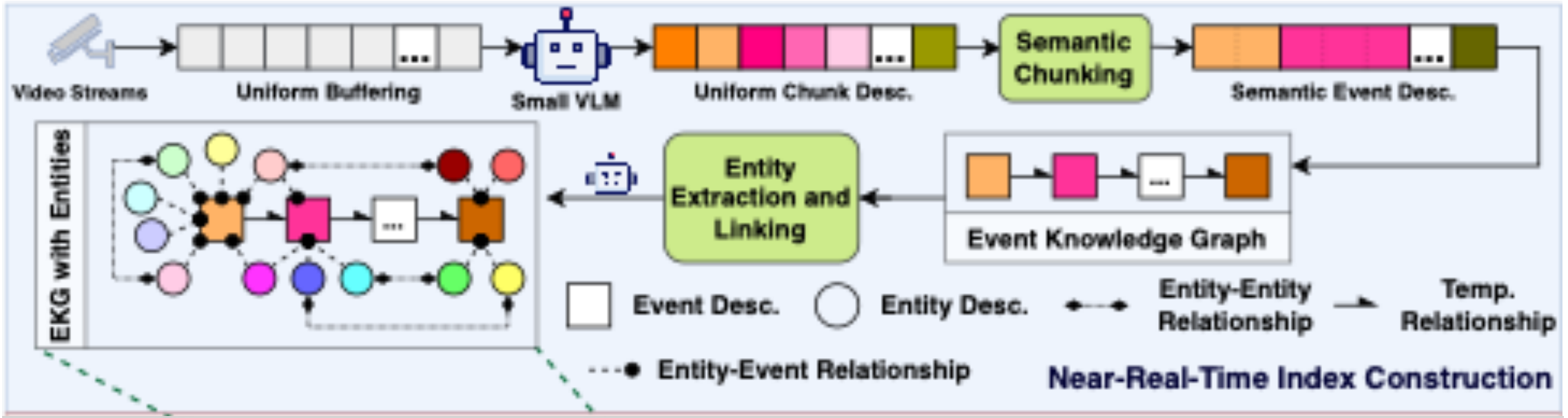
Q: At what time does the chef add the bay leaves into the pot?

4 frames

Our System Design



Our System Design



How to construct an Index Structure：

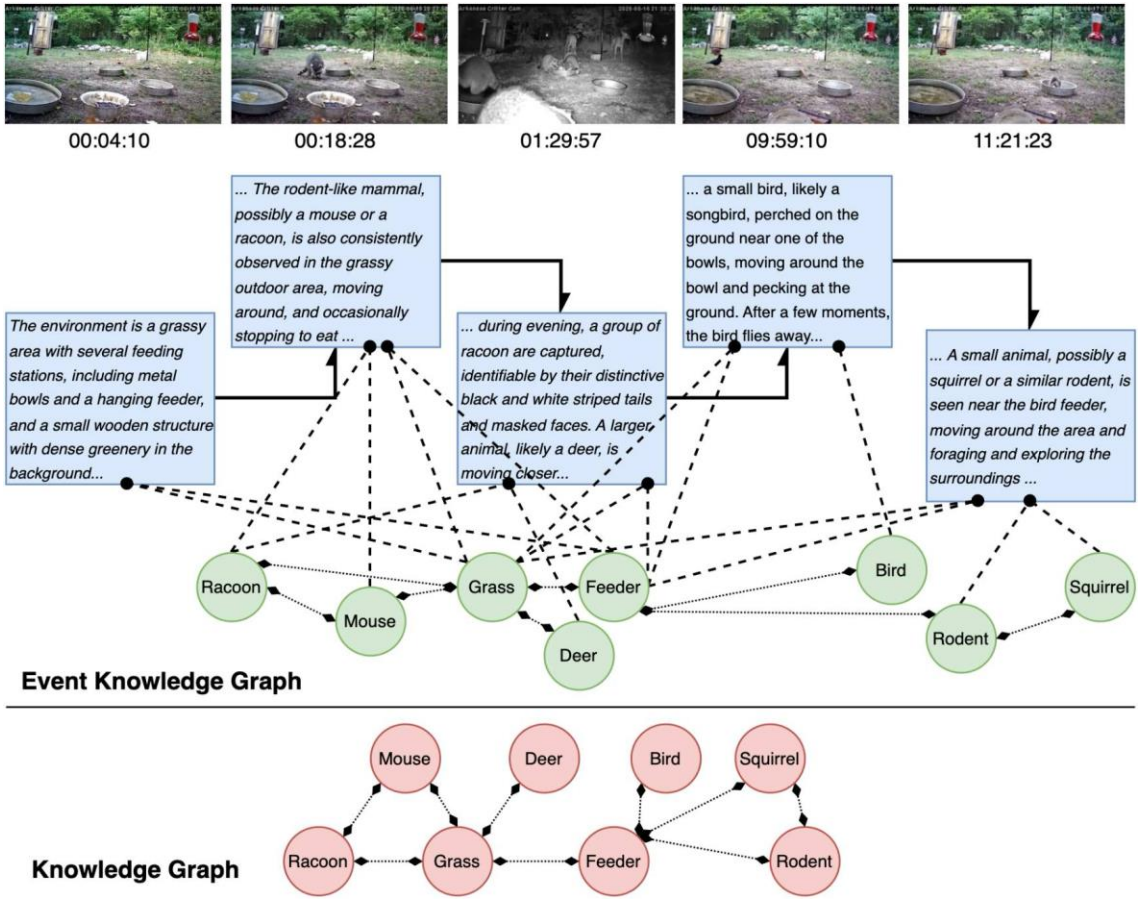
- Represents video information
- Captures complex intra-video relationships

Our System Design – Index Construction

Event Knowledge Graph

✓ **Videos are not frame sequences**

✓ **Videos are graphs of events**



Q: Why did the squirrel leave the feeder around 11:21?

1 Frame Sequence (uniform sampling)

- Only 5 sampled frames
- No temporal order
- No causal reasoning

✗ Cannot infer cause

2 Plain KG (entity only)

Squirrel — Feeder

Squirrel — Grass

...

- Entities connected
- No when / no why

✗ No when / why

3 Event KG (Ours) (events + relations)

09:59 Deer approaches

10:30 Squirrel alerts

11:21 Squirrel flees

...

✓ Causal chain recovered

1 Visual content

Q: "*What color is the bird?*"

Plain KG: ✓

Event KG: ✓

2 Casual

Q: "*Why did the squirrel leave?*"

Plain KG: ×

Event KG: ✓

3 Multi-hop

Q: " Did the deer's arrival cause the squirrel to leave?"

Plain KG: ×

Event KG: ✓

...

Our System Design – Index Construction

Video stream

.....

Timeline

Step.1 Fixed-window Captioning

Window 1 | Window 2 | Window 3 | Window 4

Desc 1
Car drives on the highway ...

Desc 2
Car changes lane ...

Desc 3
Car exits the highway ...

Desc 4
Car continues on ...

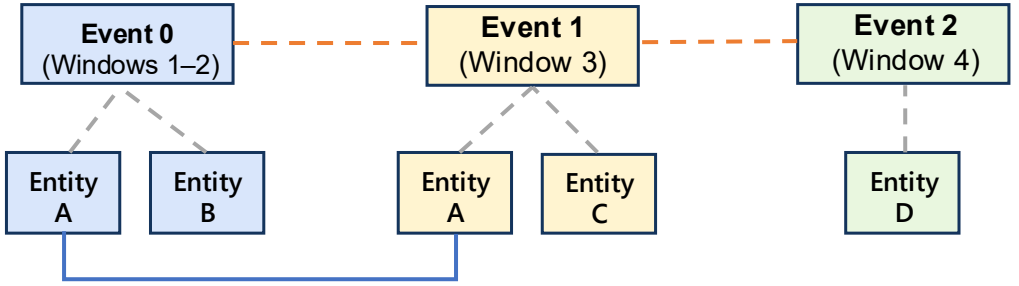
QwenVL-2.5 7B
Batch-inference

Step.2 Semantic Merging

Desc 1	Desc 2	BERT-Score(W1, W2) = 0.83	✓ merge	Event 0 (Windows 1–2)
Desc 2	Desc 3	BERT-Score(W2, W3) = 0.41	✗ split	Event 1 (Window 3)
Desc 3	Desc 4	BERT-Score(W3, W4) = 0.38	✗ split	Event 2 (Window 4)

BERT-Score

Step.3 Relation Building

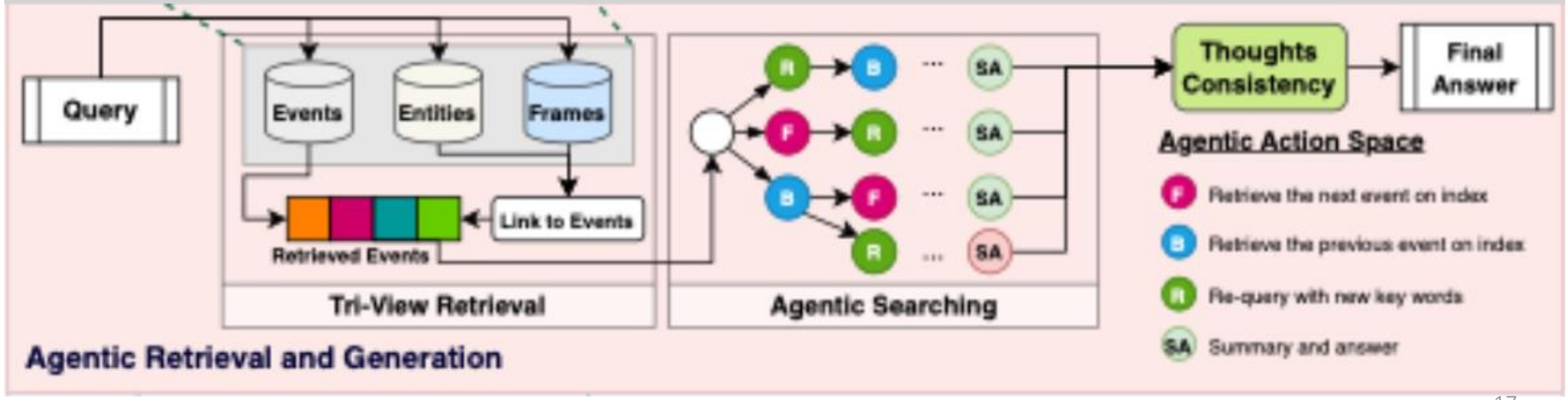


Event - Event	Temporal
Entity - Entity	Co-reference
Event - Entity	Appears-in

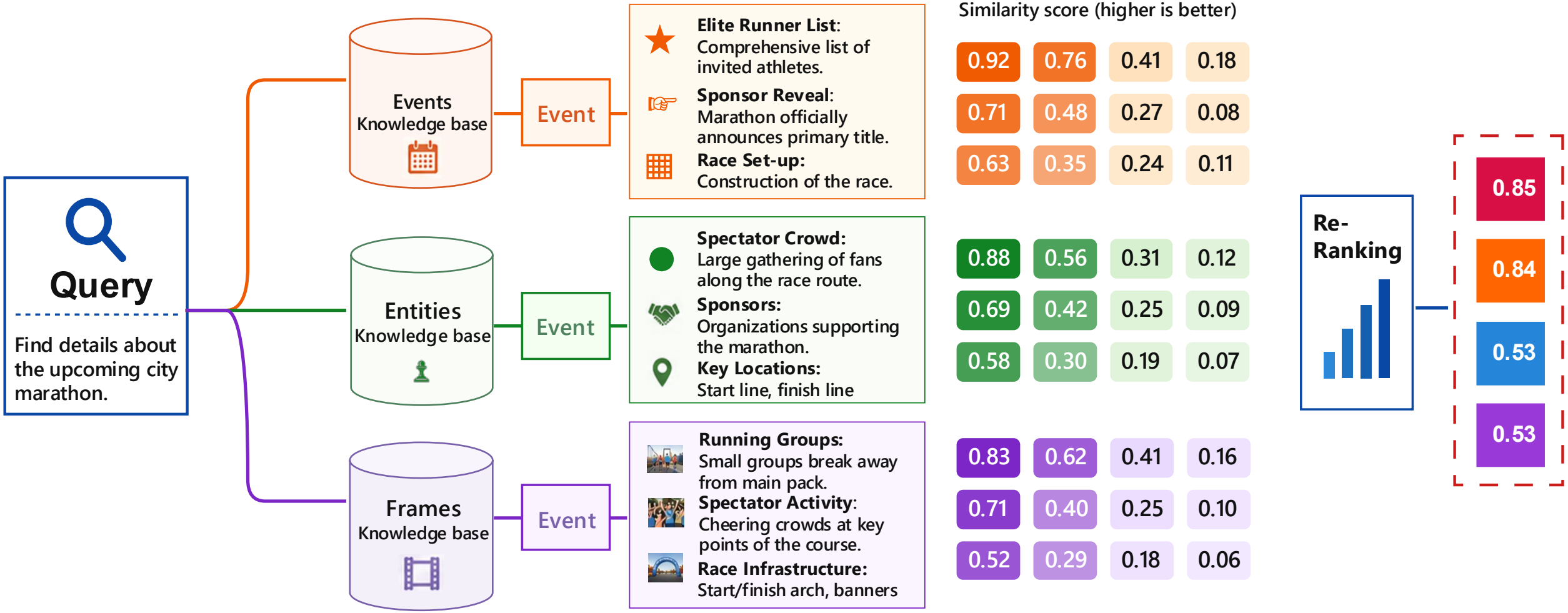
Our System Design

How to Retrieve Sufficient Information：

- Retrieves complementary evidence
- Verify answer reliability



Our System Design – Agentic Retrieval and Generation



Textual modality emphasizes **fine-grained Information**

Visual modality provides **comprehensive global information**

Our System Design – Agentic Retrieval and Generation

Tree Search

Agent explores the video by expanding actions at each step.

How Humans Watch a Video

Recap / Summarize
After watching, summarize and recall

Fast-forward
Skip ahead, see what happens

Rewind
Go back, check the details

Re-search
Not clear? Search again

Action Space

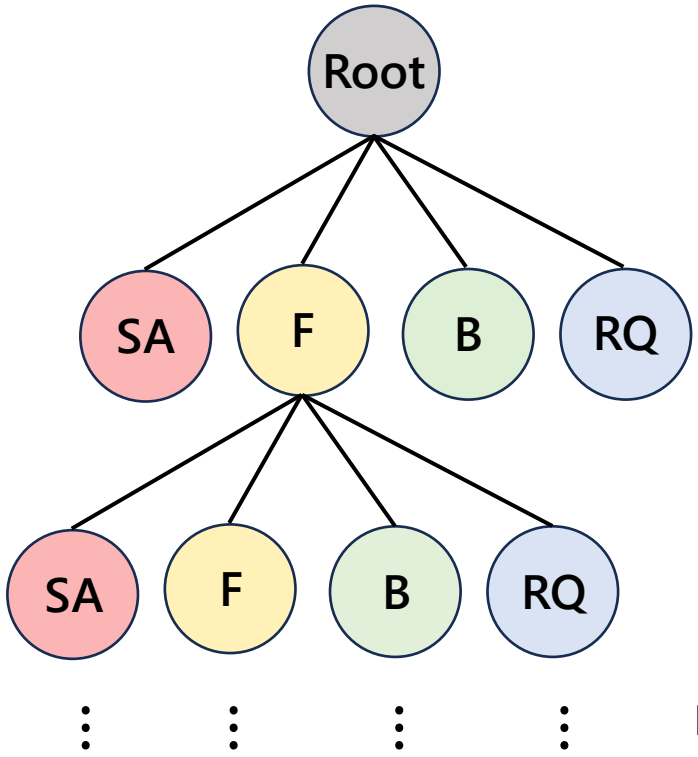
4 atomic operations the agent can take at each node

S Summary and Answer
Summarize information and generate answer

F Foward
Retrieve the following event in the timeline

B Backward
Retrieve the Preceding event in the timeline

RQ Re-query
Generate a sub-queries based on existing info



⌘The number 0f SA nodes is $\frac{4^{n+1}-4}{3}$

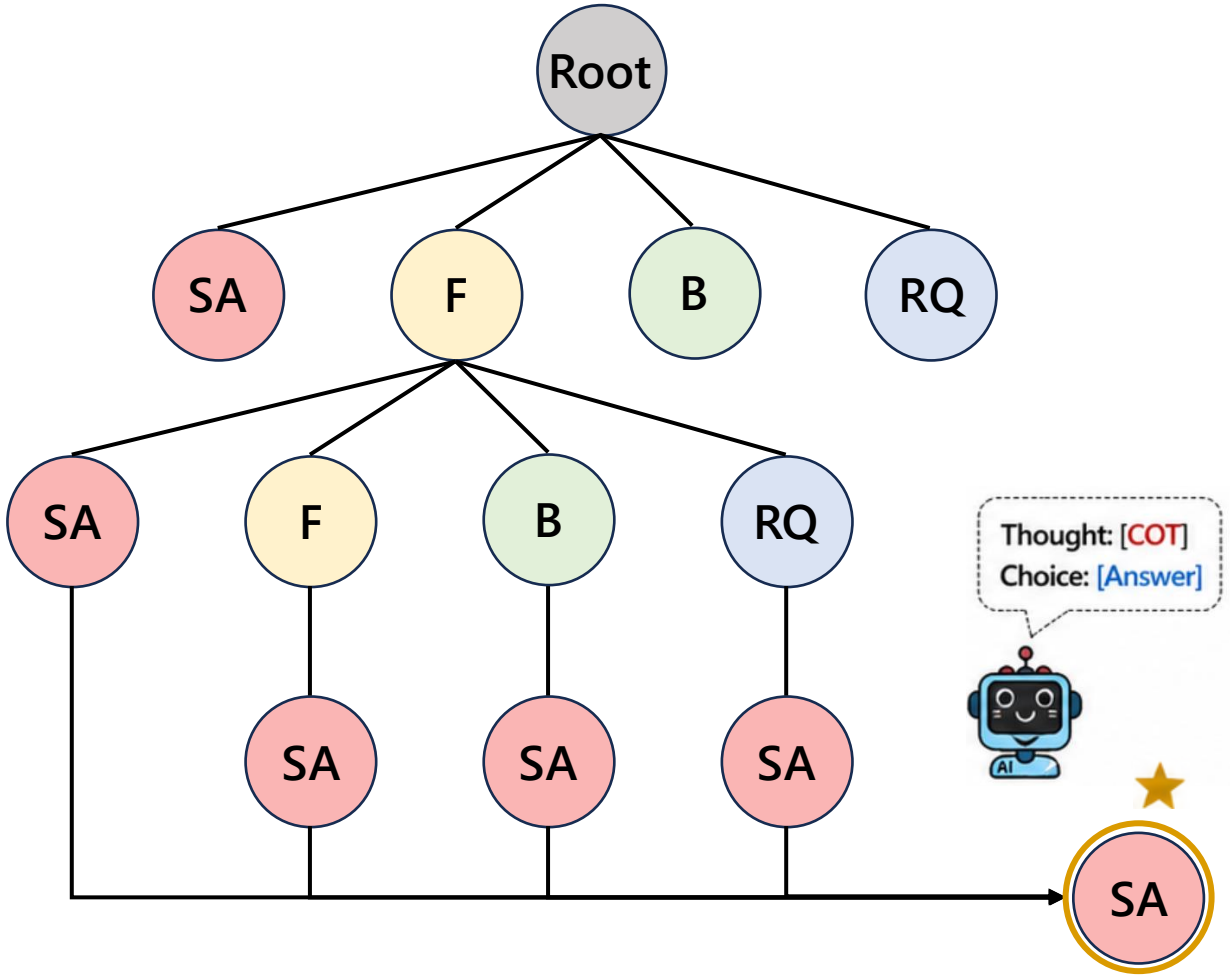
Our System Design – Agentic Retrieval and Generation

✕ Majority Vote

- Most nodes are noisy retrievals.
- The few information-rich SA nodes lose.

✓ Thought Consistency (ours)

- Score by reasoning agreement, not headcount.
- Minority but coherent → wins.



When k thoughts agree → evidence is enough

$$S_{COT}^{(t)} = \frac{2}{k(k-1)} \sum_{1 \le i < j \le k} \text{BERTScore}(COT_i, COT_j)$$

$$S_{A}^{(t)} = \frac{\left| \{ i \mid a_i = a^{(t)} \} \right|}{n}$$

$$S_{final}^{(t)} = \lambda S_{A}^{(t)} + (1-\lambda) S_{COT}^{(t)}$$

Evaluation

3 Benchmarks x 11 Baselines · Duration from 40 min → 10 hours (15x)

BENCHMARKS

VideoMME-Long [ICCV'25]

▣ 300 videos · 900 QAs

◷ ~ 40 min / video

1×

LVBench [CVPR'25]

▣ 102 videos · 1549 QAs

◷ ~ 68 min / video

1.7×

★ AVA-100

▣ 8 videos · 120 QAs

◷ ~ 10 hours / video

15×

BASELINES

VLM (Single-pass)

- GPT4-o
- Gemini-1.5-Pro
- QwenVL-2.5-7B
- InternVL-2.5-8B
- Phi-4-Multimodal
- LLaVA-Video

Sample strategy

- Uniform Sampling
- Top-K retrieval (vectorized)

Video-RAG (Retrieval-Augmented / Agentic)

- VideoTree [CVPR' 25]
- VideoAgent [ECCV' 24]
- DrVideo [CVPR' 25]
- VCA [ICCV' 25]

Evaluation

Our benchmark for hours-long, real-world video analytics

★ AVA-100	8 Videos	120 QAs	~10 h Duration	4 Scenarios

Human Daily Activities


Q: *What event follows putting oil on the surface?*

✓ **A. Spreading oil in pan**
B. Wash hands
C. Toasting bread
D. Placing plates

City Walking


Q: *Where does the person enter at the end?*

A. Flower shop
✓ **B. Creperie**
C. Bookstore
D. KFC

Wildlife Surveillance


Q: *Which animals appear in the video?*

✓ **A. Bird, Raccoon, Deer**
B. Cat, Dog, Squirrel
C. Fox, Rabbit, Bear
D. Fox, Deer

Traffic Monitoring


Q: *How many cars run the red light?*

A. 0
B. 1
✓ **C. 2**
D. 3

Evaluation – Overall Performance

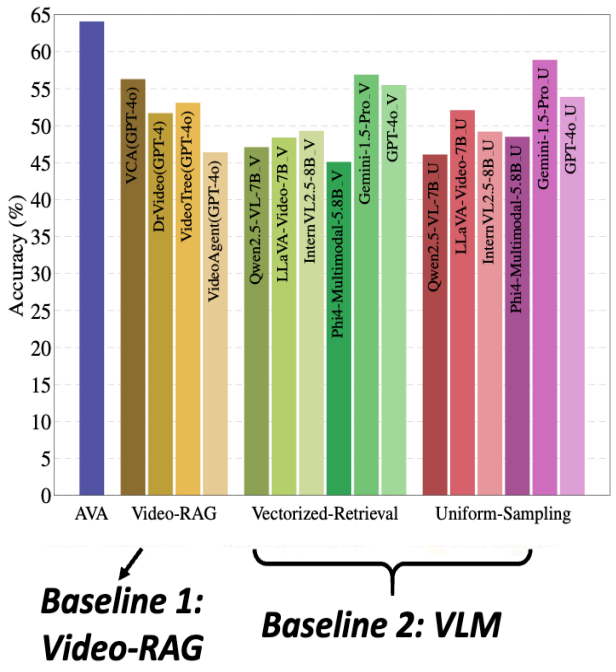
AVA consistently outperforms strong baselines

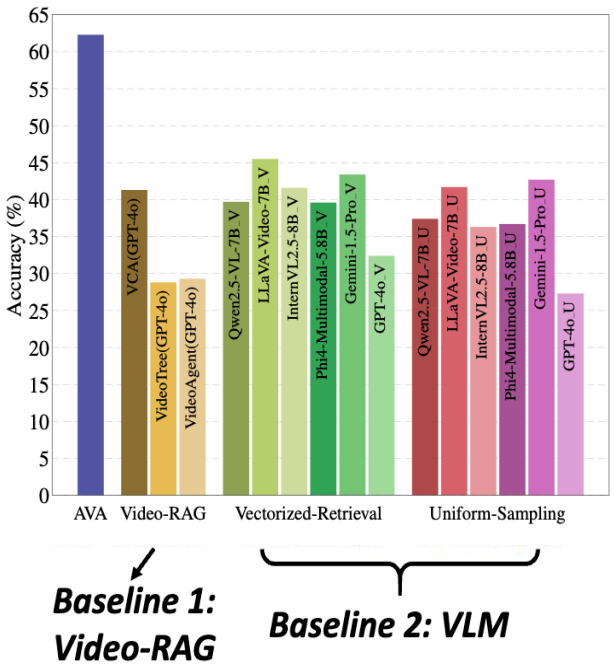
VideoMME-Long ~ 40min + 5.2% ↑ vs. best baseline

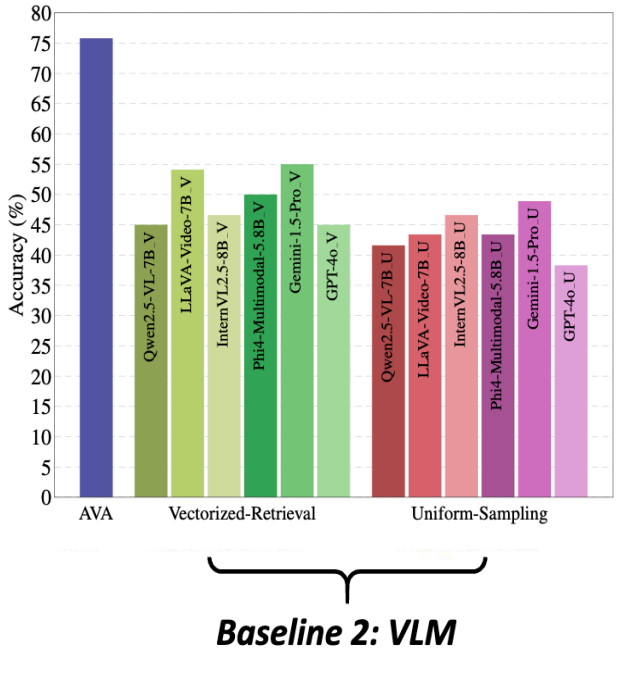
LVBench ~ 68min + 16.9% ↑ vs. best baseline

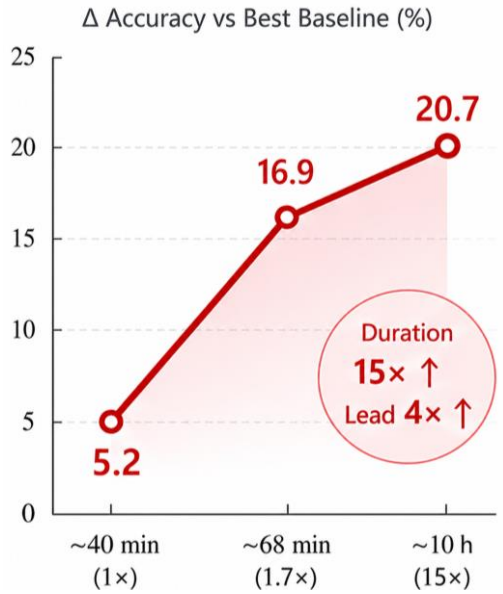
AVA-100 ~ 10h + 20.7% ↑ vs. best baseline

AVA's Lead Grows with Video length









The longer the video, the larger AVA's lead.

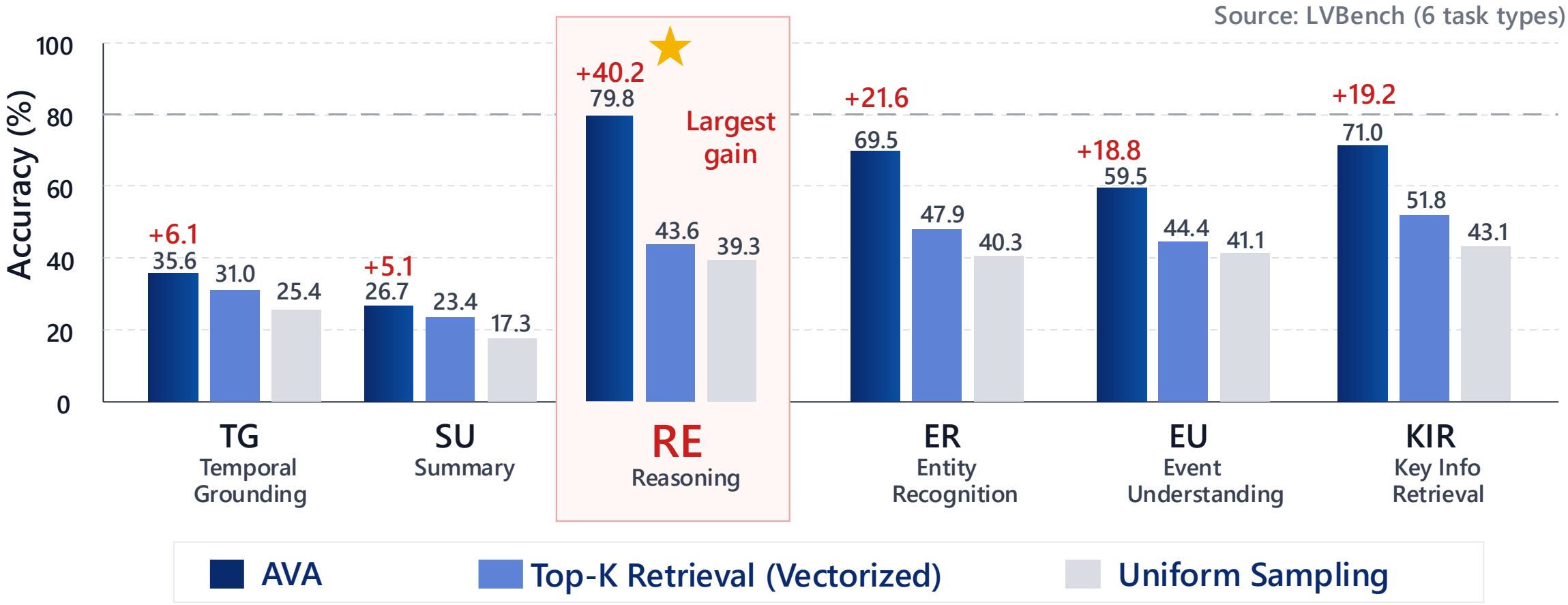
+ 5.2% ↑ → + 16.9% ↑ → + 20.7% ↑

~40min (1x) ~68min (1.7x) ~10h (15x)

Across 3 long-video benchmarks, AVA sets a new state-of-the-art and the advantage scales with video length.

Evaluation – Task Specific Performance

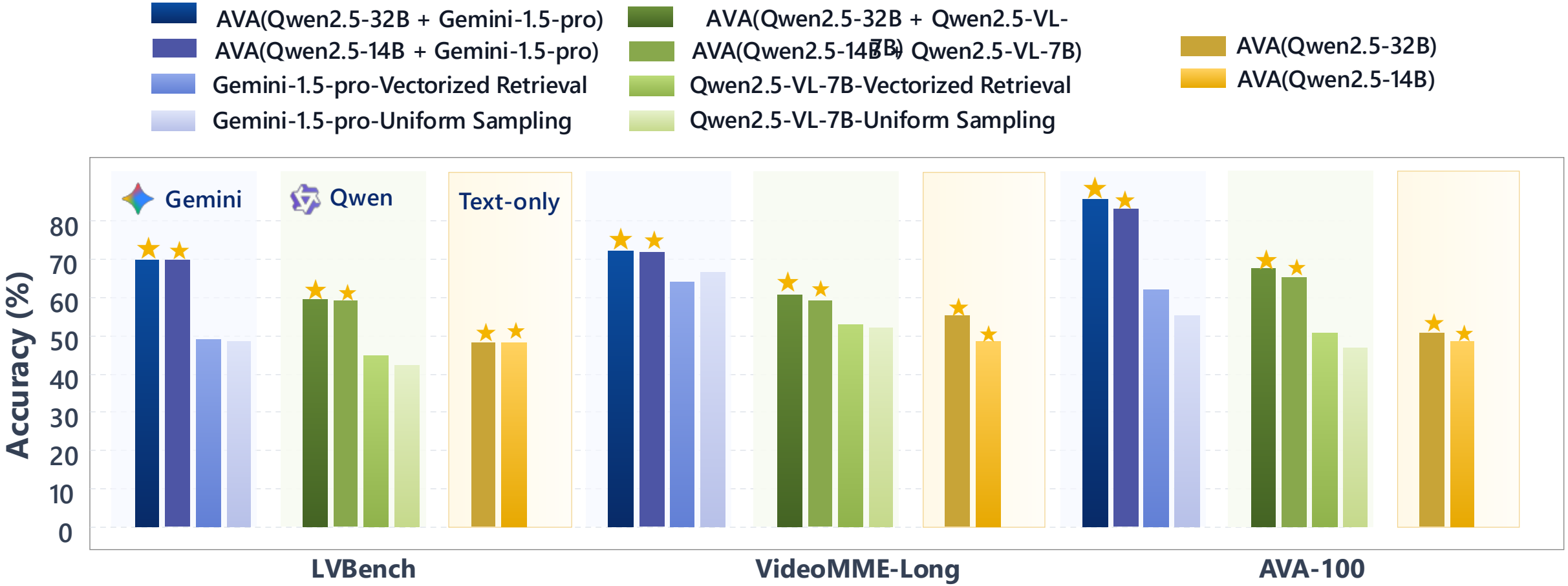
AVA wins on every task — biggest gain on Reasoning.



Reasoning (RE) sees the largest gain - **+40.2** vs the best baseline - L4 core ability

Evaluation – Configurations

AVA with different configurations.



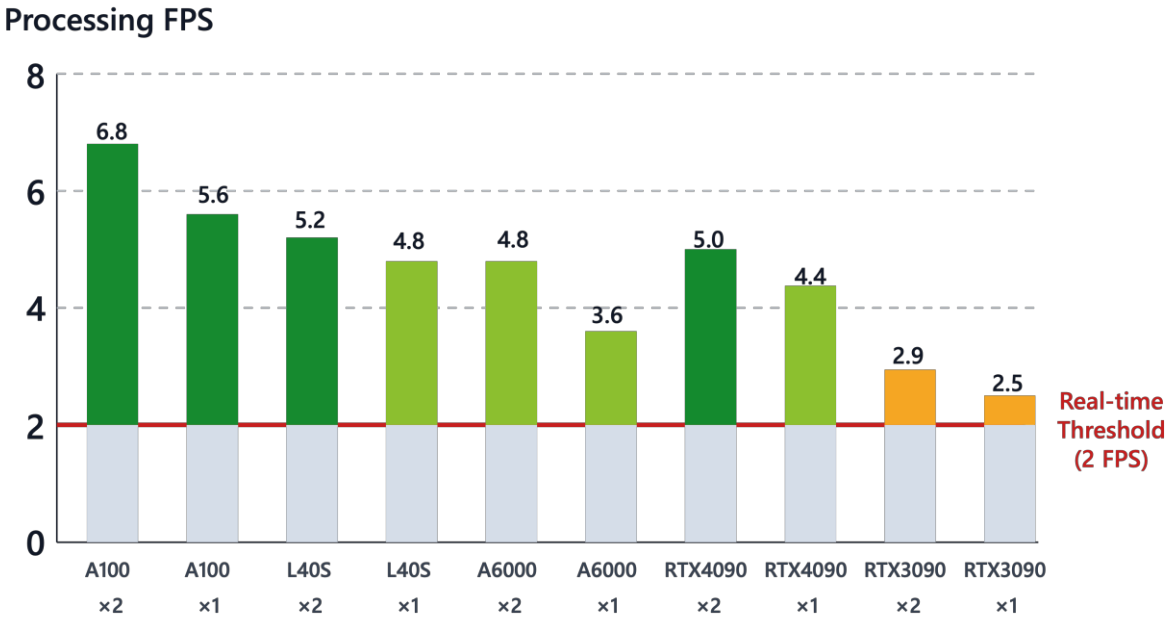
AVA configurations are emphasized across all benchmark groups.

AVA with only text (highlighted in yellow) can also outperform many baselines.

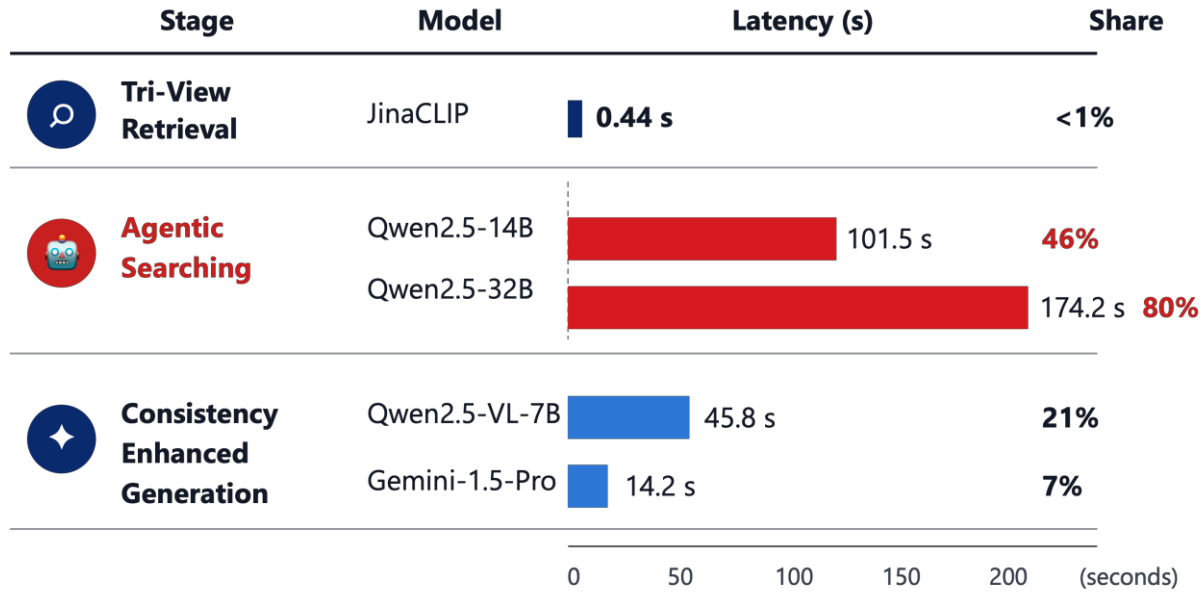
Evaluation – Overhead

Index runs in real-time. Inference is LLM-bound.

Index Construction Time



Search & Generation Time



Index: real-time on every GPU

All GPU configurations achieve ≥ 2 FPS, keeping construction in real time.

Inference: agentic loop is the bottleneck.

Agentic searching takes ~80% of total latency.

Conclusion & Future Direction

AVA pushes video analytics toward L4 – L5.

Conclusion

1. **AVA**: the first L4 video analytics system powered by VLMs
2. **AVA-100**: a 10-hour benchmark for real-world video analytics
3. +20.8% QA Acc. over the best baselines

Future Direction

1	**Scale**: AVA-100 is still limited in size	**Scale AVA-100:** more videos, domains, and richer task types
2	**Overhead**: agentic rollout expands all nodes	**Learned action policy:** train to decide the next action at each step
3	**VLM weakness**: some visual skills remain unreliable, e.g., counting	**Expert models:** call specialized visual experts and move toward L5 analytics

Keep moving towards L5 Video Analytics

AVA

Towards Agentic Video Analytics with Vision Language Models

Code

Benchmark

USENIX

THE ADVANCED COMPUTING SYSTEMS ASSOCIATION

AVA: Towards Agentic Video Analytics with Vision Language Models

Yuxuan Yan, *Zhejiang University;* Shiqi Jiang, *Microsoft Research;* Ting Cao, *Tsinghua University;* Yifan Yang, *Microsoft Research;* Qianqian Yang and Yuanchao Shu, *Zhejiang University;* Yuqing Yang and Lili Qiu, *Microsoft Research*

https://www.usenix.org/conference/nsdi26/presentation/yan

This paper is included in the Proceedings of the 23rd USENIX Symposium on Networked Systems Design and Implementation.

May 4–6, 2026 • Renton, WA, USA

ISBN 978-1-939133-54-0

Open access to the Proceedings of the 23rd USENIX Symposium on Networked Systems Design and Implementation is sponsored by