SOUTHEAST UNIVERSITY 東南大學

计算机科学与工程学院
School of computer science and engineering

[NSDI'26] 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*

https://github. com/I-ESC/Project-Ava

Presenter：Tianen Liu

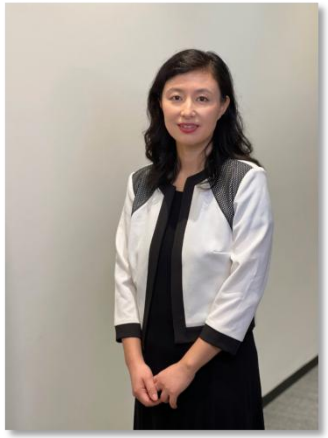
11/17/2025

Team

Shiqi Jiang
MSRA

Yuanchao Shu
Zhejiang University

Ting Cao
Tsinghua University



Lili Qiu
MSRA

Topic:

- Edge AI/ML analytics
- AI inference systems
- Mobile systems and AIoT

Top-tier computer system conferences:

- ISCA, ASPLOS, MobiCom, MobiSys, NSDI, OSDI, PLDI, EuroSys, SC, and PPoPP

- [NSDI'24] Vulcan: Automatic Query Planning for Live ML Analytics
- [NSDI'23] GEMEL: Model Merging for Memory-Efficient, Real-Time Video Analytics at the Edge
- [NSDI'23] RECL: Responsive Resource-Efficient Continuous Learning for Video Analytics
- [EuroSys'26] Scaling LLM Test-Time Compute with Mobile NPU on Smartphones
- [MobiCom'25] Confidant: Customizing Transformer-based LLMs via Collaborative Training on Mobile Devices

Outline

1 Background

2 Motivation

3 Design

4 Evaluation

5 Conclusion

Outline

1 Background
2 Motivation
3 Design
4 Evaluation
5 Conclusion

Background

- **Video analytics**

Object detection task

DNN-based model, e.g., YoLo

Video question answering task

Qusetion: *Why does the train stop for a while before moving again at the end of the video ?*

LLM/VLM

Answer: *The train stops for a scheduled stop at a station*

Video query task

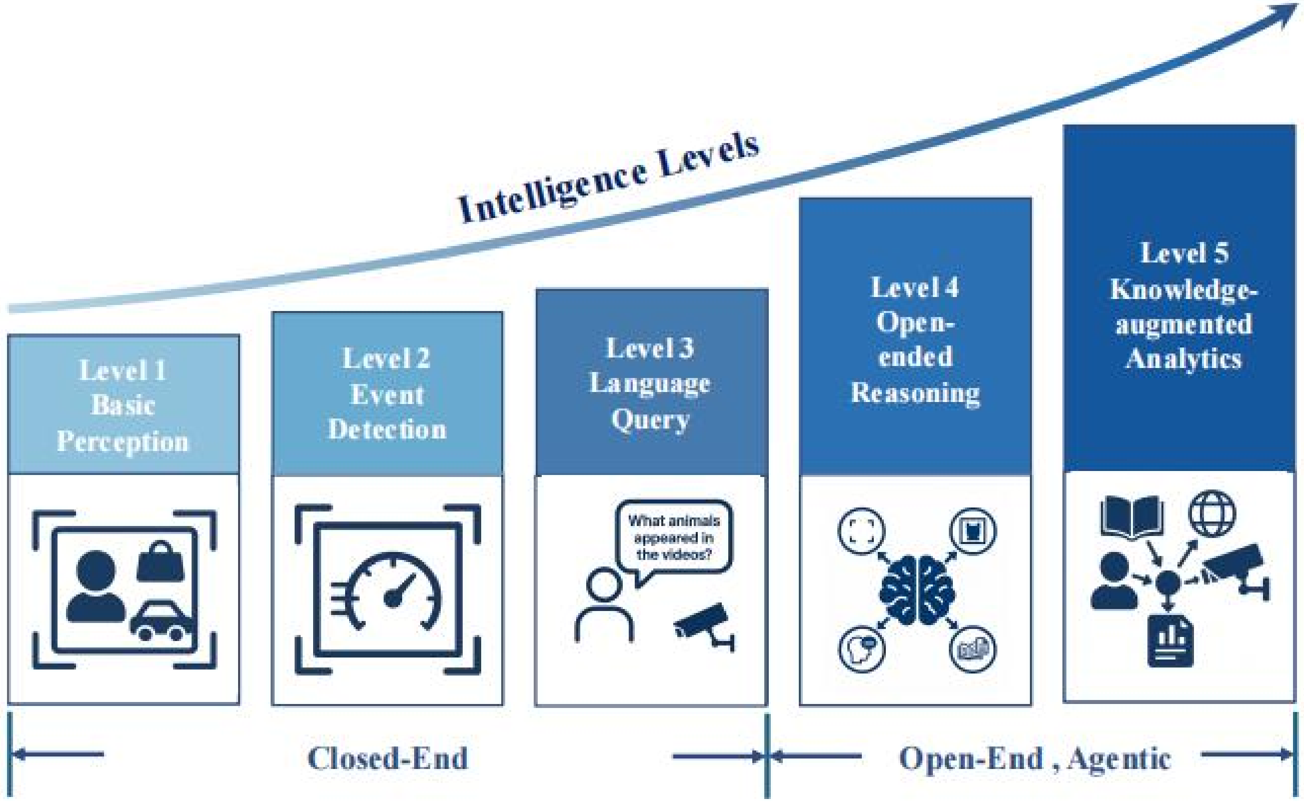
Query: *Can you confirm if the vehicle the uploaded image has been witnessed in Beijing?*

LLM/VLM

Knowledge graph

Answer: *Yes*

Background



Intelligence levels of video analytics systems

Closed-End video analytics system:

- **L1 ~ L3**
- **Predefined task/query, e.g., detection**
- **Domain-specific model, e.g., DNN**

CAT

Detection: YoLo

Level 1

Event detection: ActionFormer

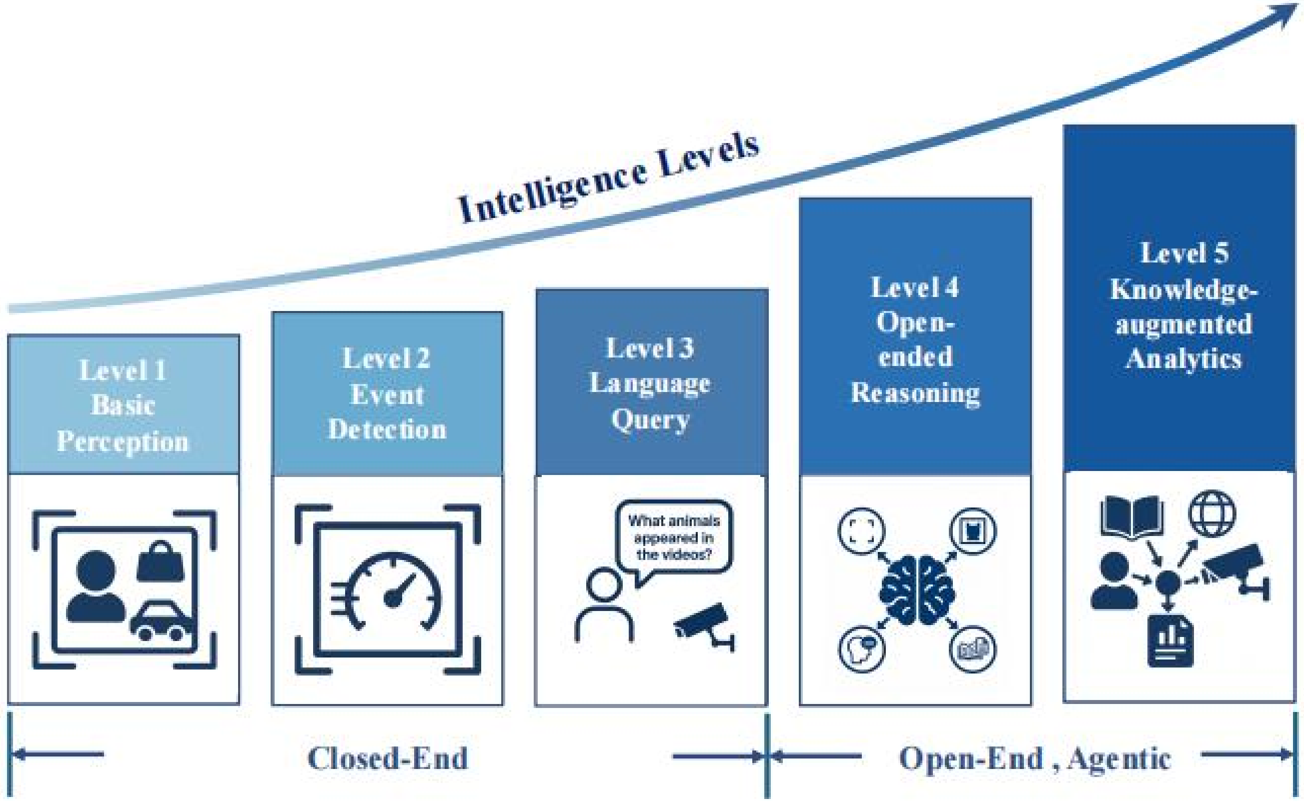
Level 2

Video query: CLIPBERT model

Level 3

Adaptability & Flexibility ✗

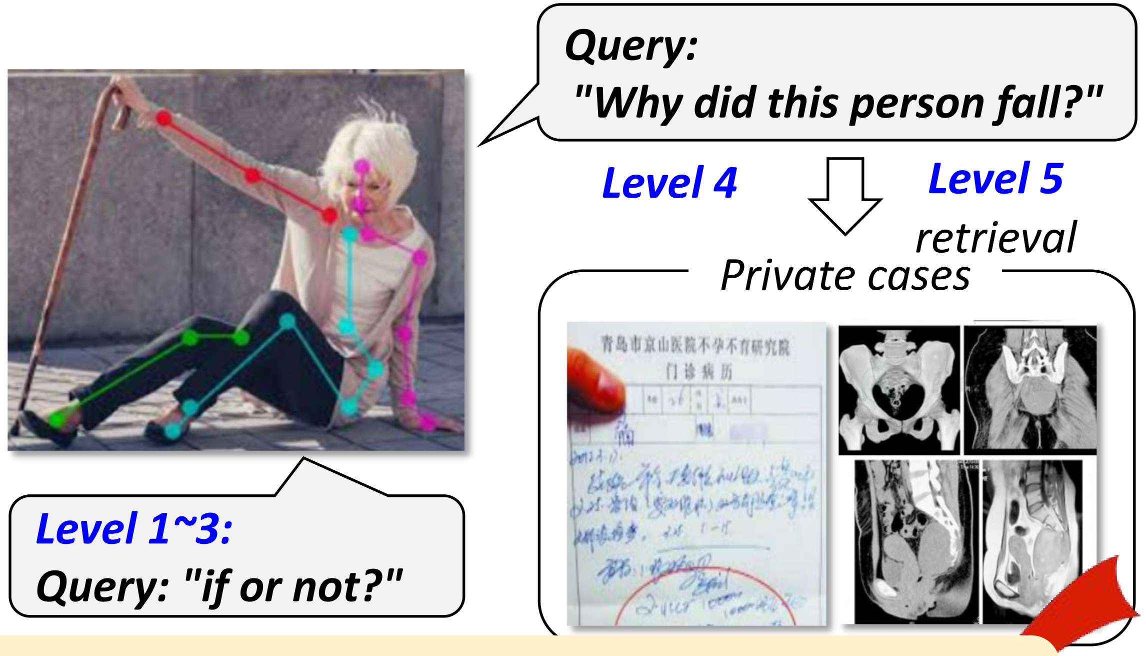
Background



Intelligence levels of video analytics systems

Open-end, Agentic video analytics system:

- L4 ~ L5
- Open and diverse tasks
- General model, e.g., LLM/VLM
- Complex query



Problem: how to achieve accurate and efficient video analytics for various tasks?

Outline

1 Background

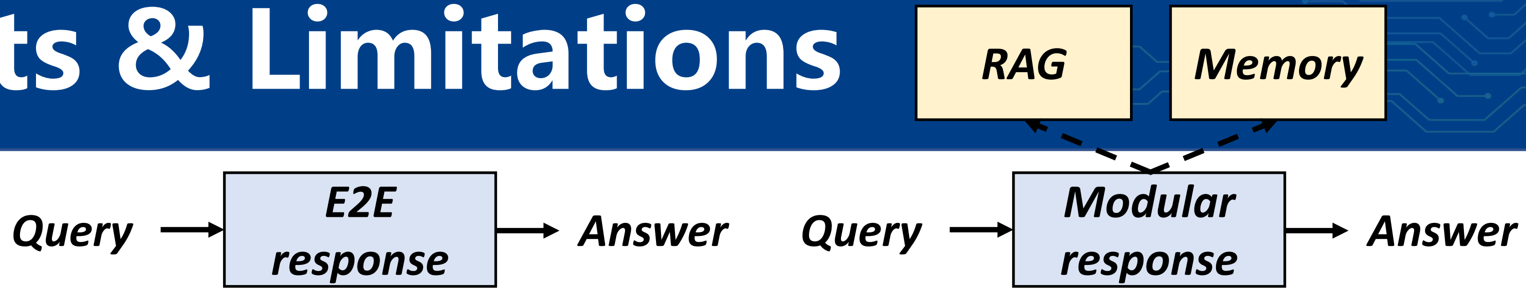
2 Motivation

3 Design

4 Evaluation

5 Conclusion

State-of-the-Arts & Limitations



	E2E response	Modular response
Closed-end analytics (L1~L3)	Vulcan [NSDI'24], RECL [NSDI'23], Gemel [NSDI'23], Ekya [NSDI'22]	Video-RAG [NeurIPS'25], VideoTree [CVPR'25], VideoAgent [ECCV'24], DrVideo [CVPR'25]
Open-end analytics (L4~L5)	VLMs like GPT-4o, Gemini, QwenVL and Phi	**[NSDI'26] AVA**

Limitation 1: Struggle to handle ultra-long videos (> 10 hours).

- L1 ~ L3: Rely on DNNs and process each video frame independently.
- L4 ~ L5: Traditional VLM limited inherent context window

Limitation 2: Struggle to handle open-end complex tasks.

- Predefined tasks, e.g., detection —> The limited agentic reasoning capabilities

Opportunity & Challenge

Limitation 1: Struggle to handle ultra-long videos.

Opportunity 1: Only a small portion of the frames are necessary to answer

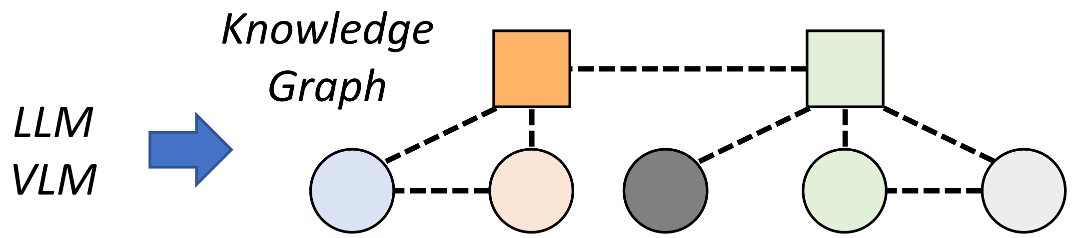
Short (1.4 minutes)		Medium (9.7 minutes)		Long (39.7 minutes)	
Total	Needed	Total	Needed	Total	Needed
2144.8	12.1 (0.5%)	13924.1	68.1 (0.4%)	66847.1	82.3 (0.1%)

Empolyed VideoMME benchmark and Qwen2-VL.

Challenge 1: How to extract useful information from ultra-long videos?

Limitation 2: Struggle to handle open-end tasks.

Opportunity 2: The LLM/VLM + knowledge graphs enables answering open-end questions.



Challenge 2: How to achieve accurate and efficient agentic searching on graph?

Outline

1 Background

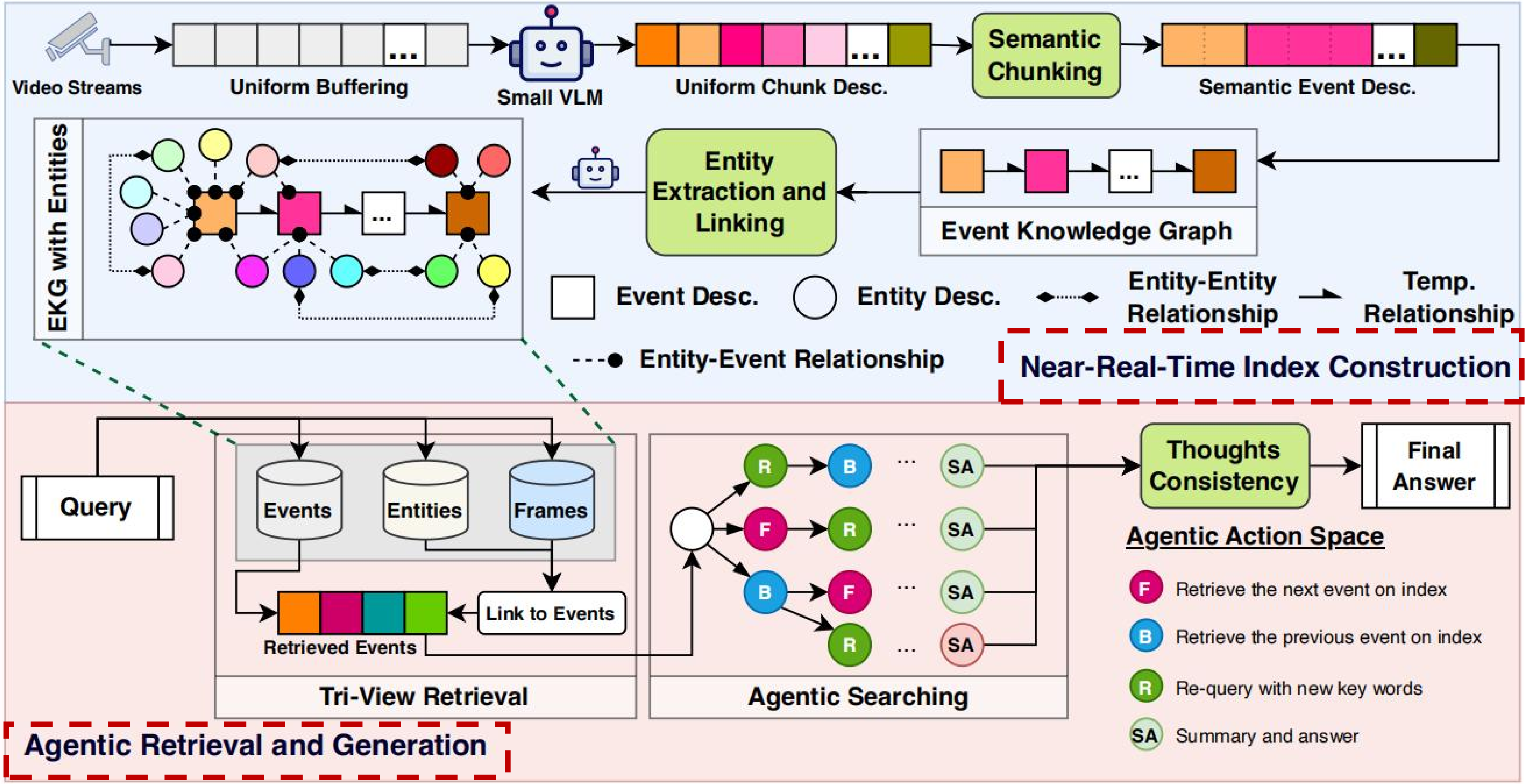
2 Motivation

3 Design

4 Evaluation

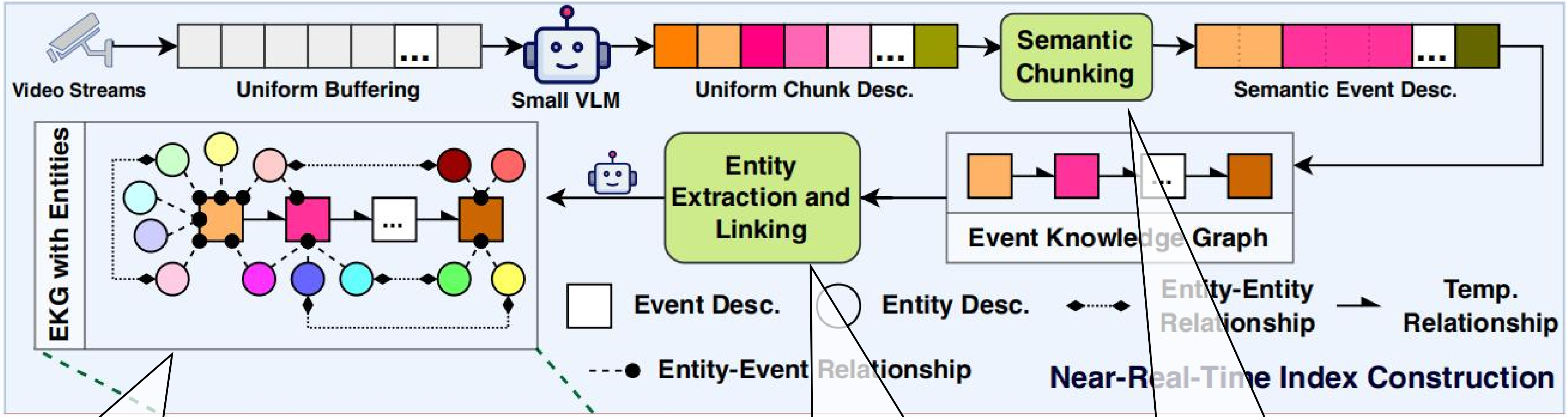
5 Conclusion

Overview



Video Streams
Uniform Buffering
Small VLM
Uniform Chunk Desc.
Semantic Chunking
Semantic Event Desc.
EKG with Entities
Entity Extraction and Linking
Event Knowledge Graph
Event Desc.
Entity Desc.
Entity-Entity Relationship
Temp. Relationship
Entity-Event Relationship
Near-Real-Time Index Construction
Query
Events
Entities
Frames
Link to Events
Retrieved Events
Tri-View Retrieval
R
B
F
SA
Agentic Searching
Thoughts Consistency
Final Answer
Agentic Action Space
F Retrieve the next event on index
B Retrieve the previous event on index
R Re-query with new key words
SA Summary and answer
Agentic Retrieval and Generation

Design 1: Near-Real-Time Index Construction



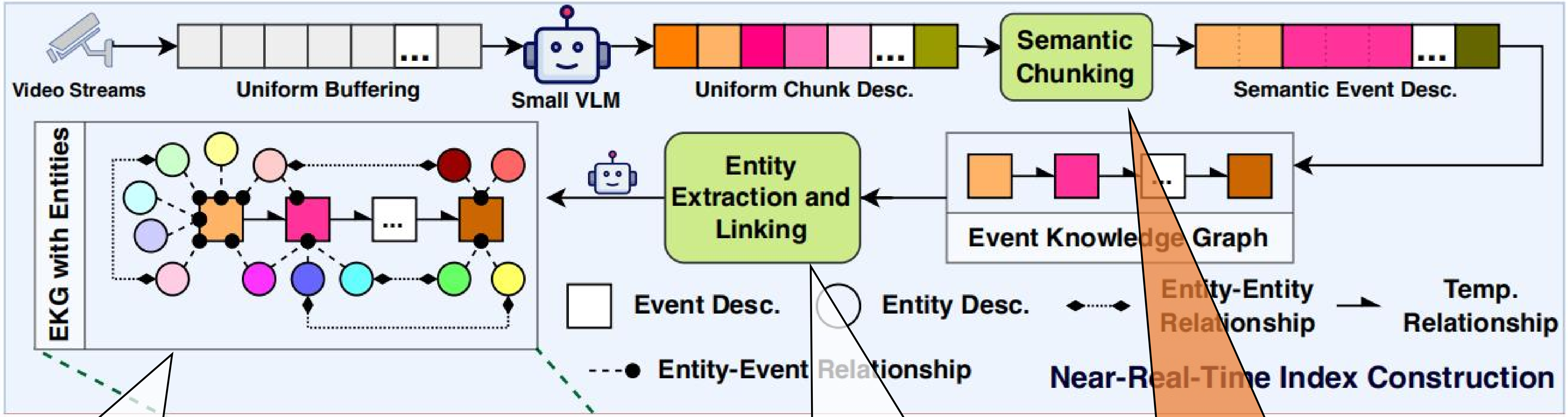
Model 3: Event Knowledge Graph with Entities

Model 2: Entity Extraction and Linking

Model 1: Semantic Chunking

Challenge 1: How to extract useful information from ultra-long videos?

Design 1: Near-Real-Time Index Construction



Model 3: Event Knowledge Graph with Entities

Model 2: Entity Extraction and Linking

Model 1: Semantic Chunking

Challenge 1: How to extract useful information from ultra-long videos?

Model 1: Semantic Chunking

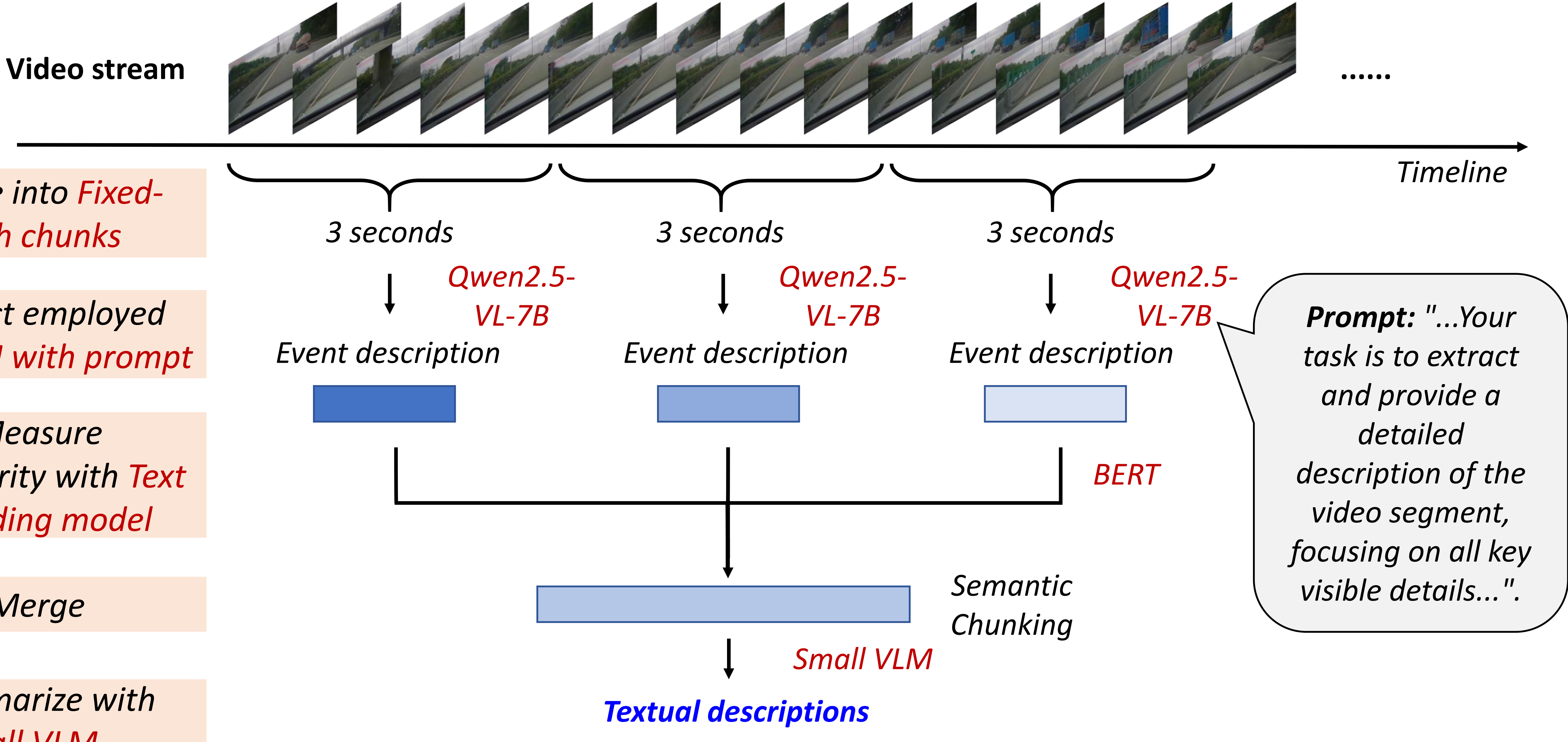
1. Divide into Fixed-length chunks

2. Extract employed Small VLM with prompt

3. Measure the similarity with Text embedding model

4. Merge

5. Summarize with Small VLM



Model 1: Semantic Chunking

Uniform Chunk Index

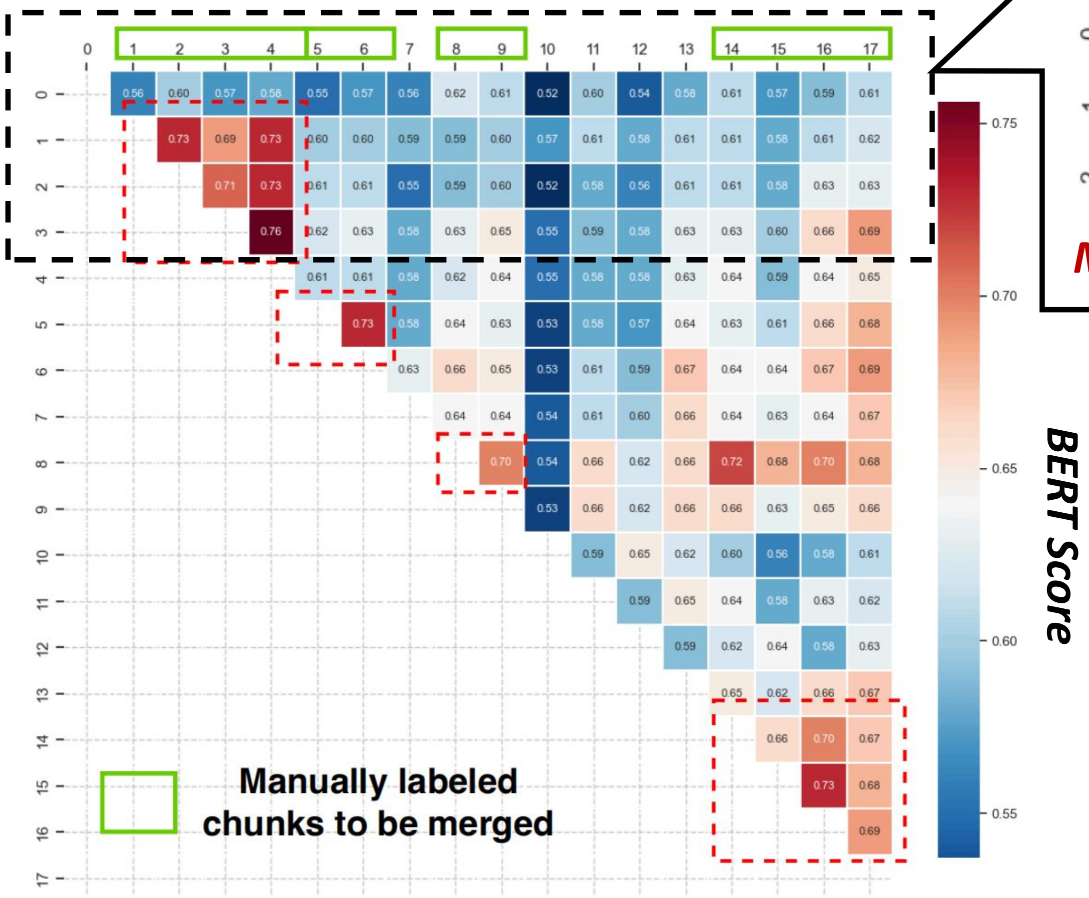
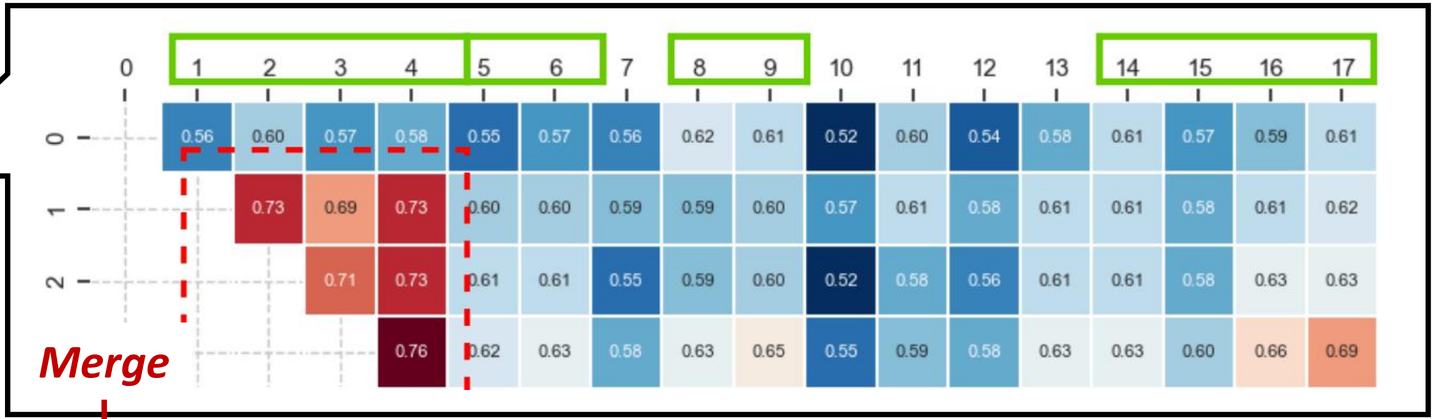


Figure: Merging uniform chunks into semantic chunks guided by the pairwise BERTScore distribution.

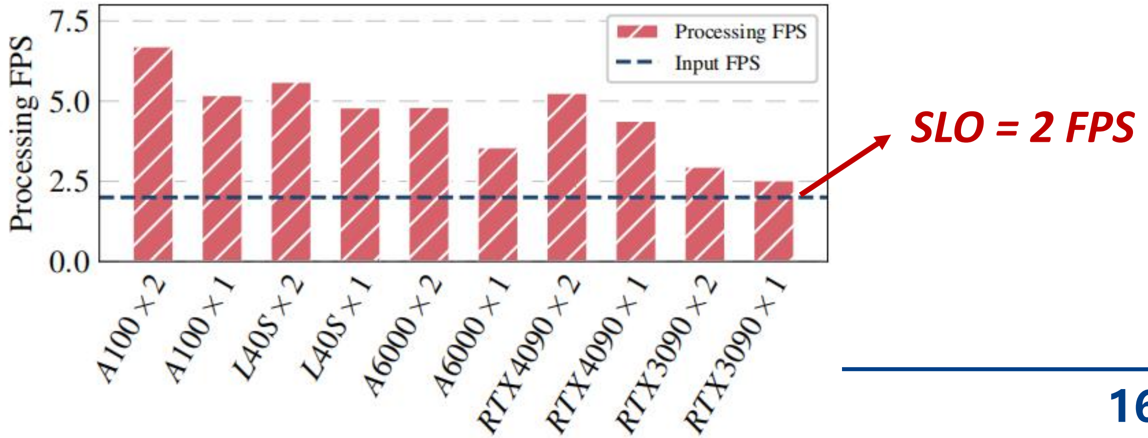


> predefined threshold → Merge 1, 2, 3, 4

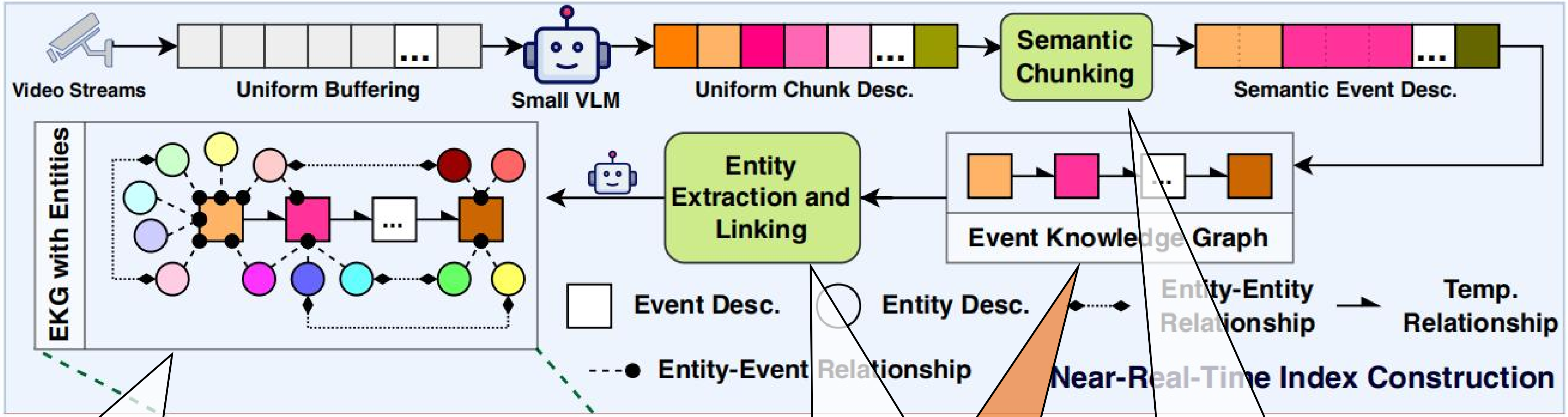
< predefined threshold → Not merge 4, 5

"......The semantic chunking process does not become a bottleneck in the near-real-time index construction phase."

- *Batch inference, maximize GPU utilization*



Design 1: Near-Real-Time Index Construction



Model 3: Event Knowledge Graph with Entities

Model 2: Entity Extraction and Linking

Model 1: Semantic Chunking

Event Knowledge Graph

...

An Example of Event Knowledge Graph

Semantic Chunk

00:04:10

Event 1

The environment is a grassy area with several feeding stations, including metalbowls and a hanging feeder,and a small wooden structure with dense greenery in the background......

Semantic Chunk

00:18:28

Event 2

... The rodent-like mammal, possibly a mouse or aracoon, is also consistently observed in the grassy outdoor area, moving around, and occasionally stopping to eat ..

Semantic Chunk

01:29:57

Event 3

.. during evening, a group ofracoon are captured, identifiable by their distinctive black and white striped tailsamd masked faces. A largeraninial likely a deer, ismoving closer..

Semantic Chunk

09:59:10

Event 4

... a small bird, likely a songbird, perched on theground near one of thebowls, moving around thebowl and pecking at theground. After a few momentsthe bird flies awav.

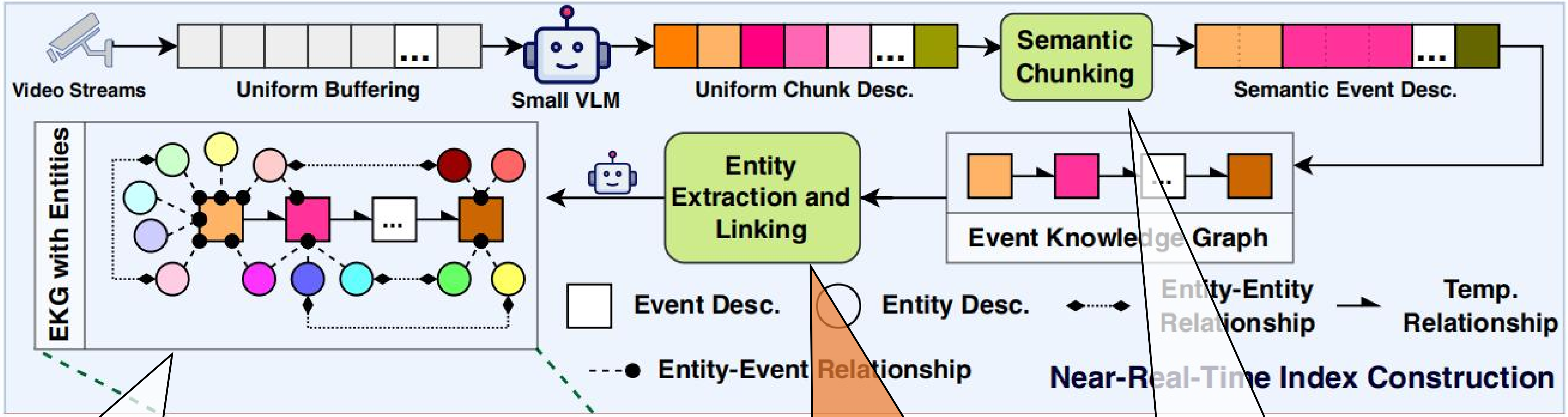
Semantic Chunk

11:21:23

Event 5

.. A small animal, possibly a squirrel or a similar rodent, is seen near the bird feeder, moving around the area and for aging and exploring the surroundings ...

Design 1: Near-Real-Time Index Construction



Model 3: Event Knowledge Graph with Entities

Model 2: Entity Extraction and Linking

Model 1: Semantic Chunking

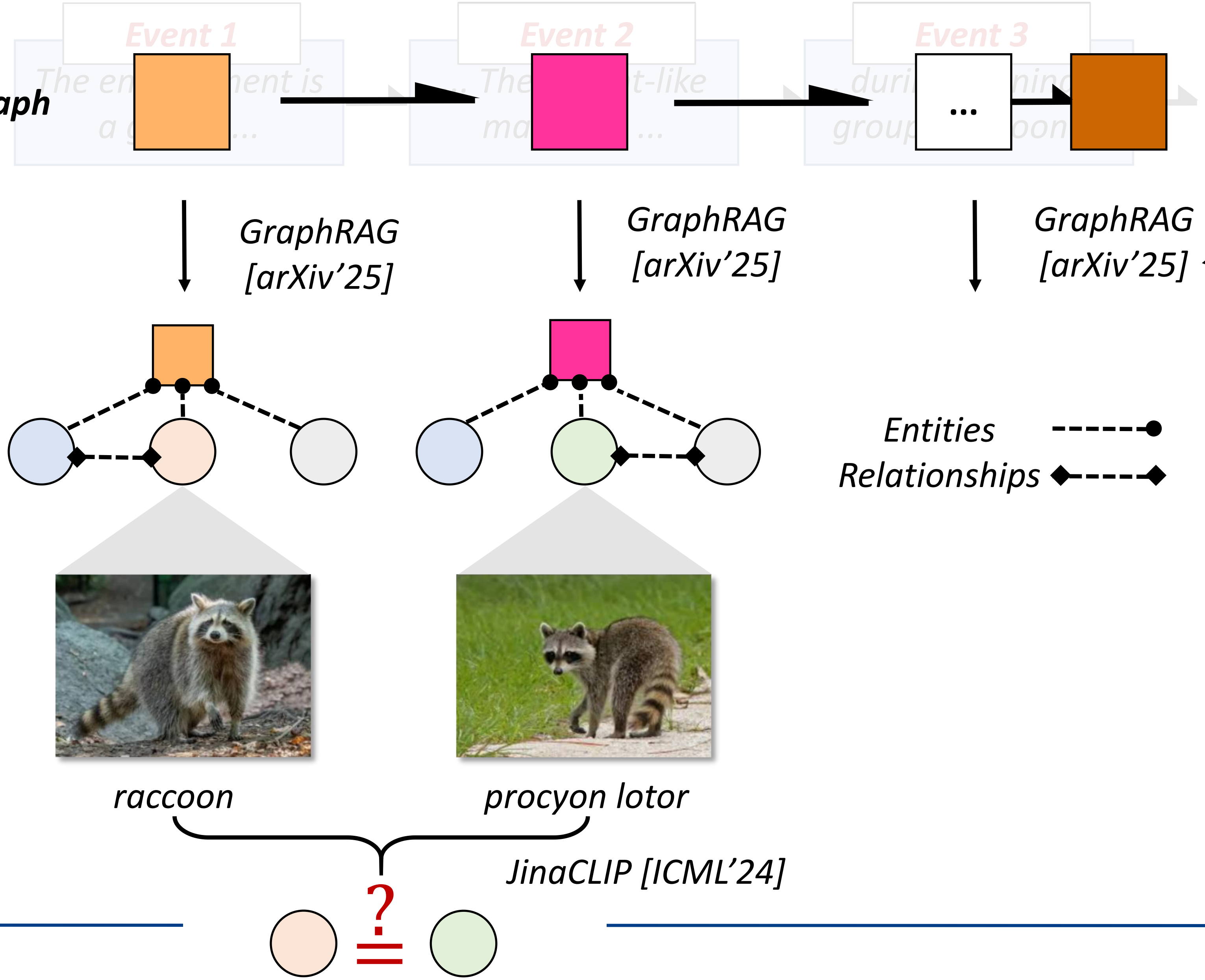
Challenge 1: How to extract useful information from ultra-long videos?

Model 2: Entity Extraction and Linking

Event knowledge graph

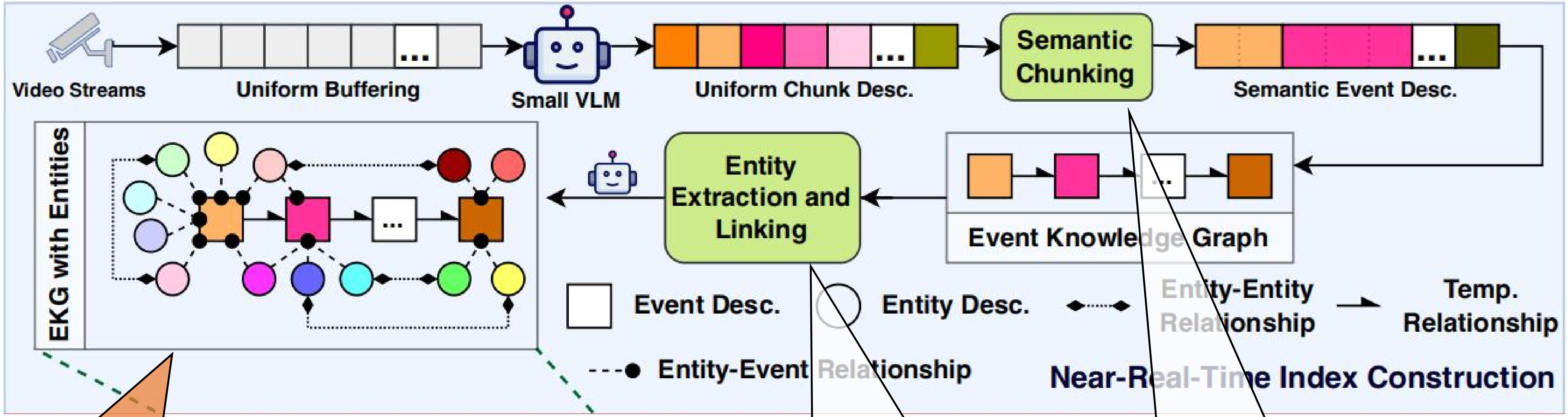
1. *Extract entity and relationships empolyed Small VLM with Prompt*

2. *De-duplicate and link employed Text Embedding Model*



***Prompt:** "...Your task is to extract and provide a detailed description of the video segment, focusing on all key visible details...".*

Design 1: Near-Real-Time Index Construction



Model 3: Event Knowledge Graph with Entities

Model 2: Entity Extraction and Linking

Model 1: Semantic Chunking

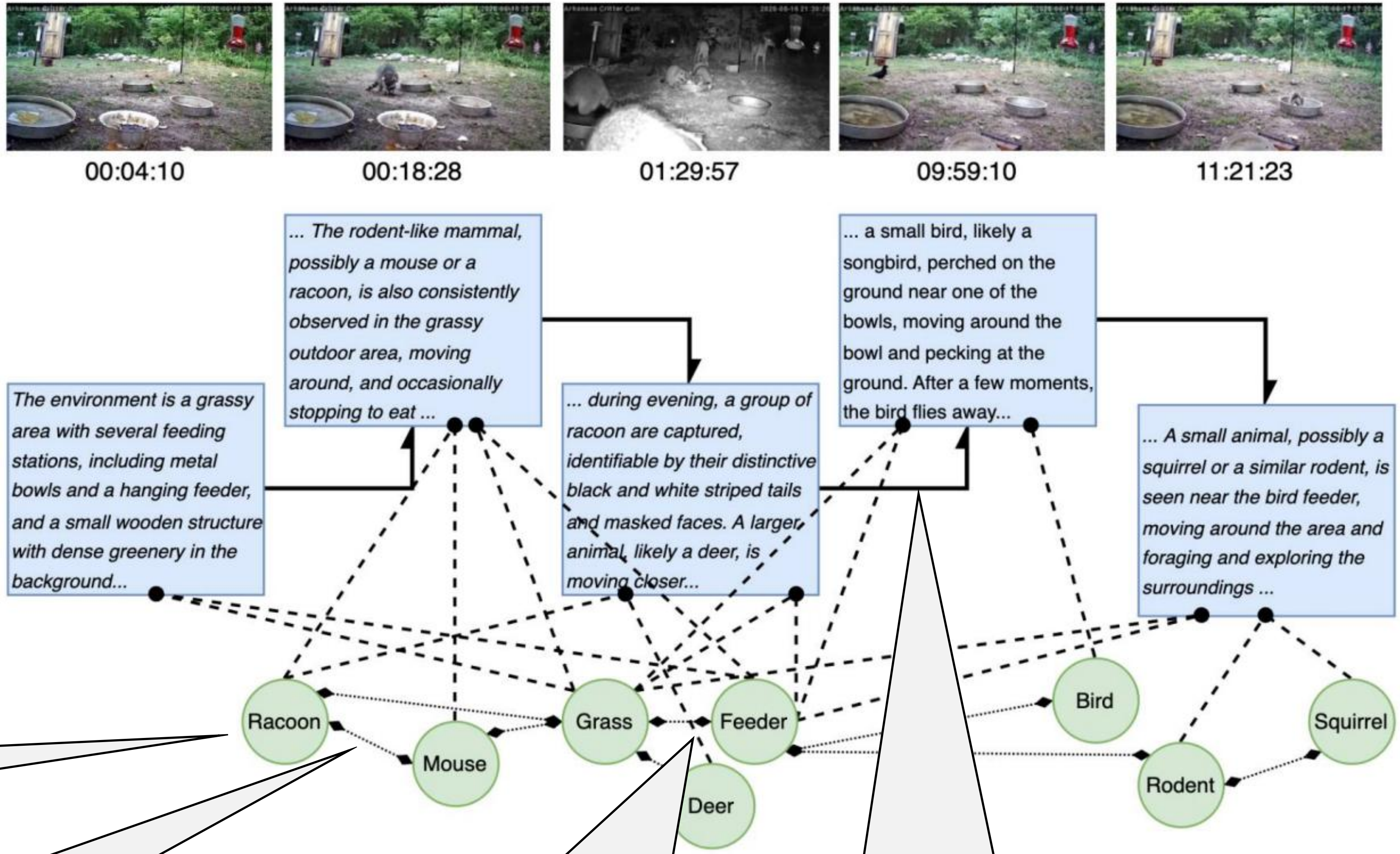
Challenge 1: How to extract useful information from ultra-long videos?

Model 3: Event Knowledge Graph with Entities

Wildlife monitoring video

Model 1

Model 2



Event Knowledge Graph

The temporally ordered set of events

The entities extracted from the video within each event

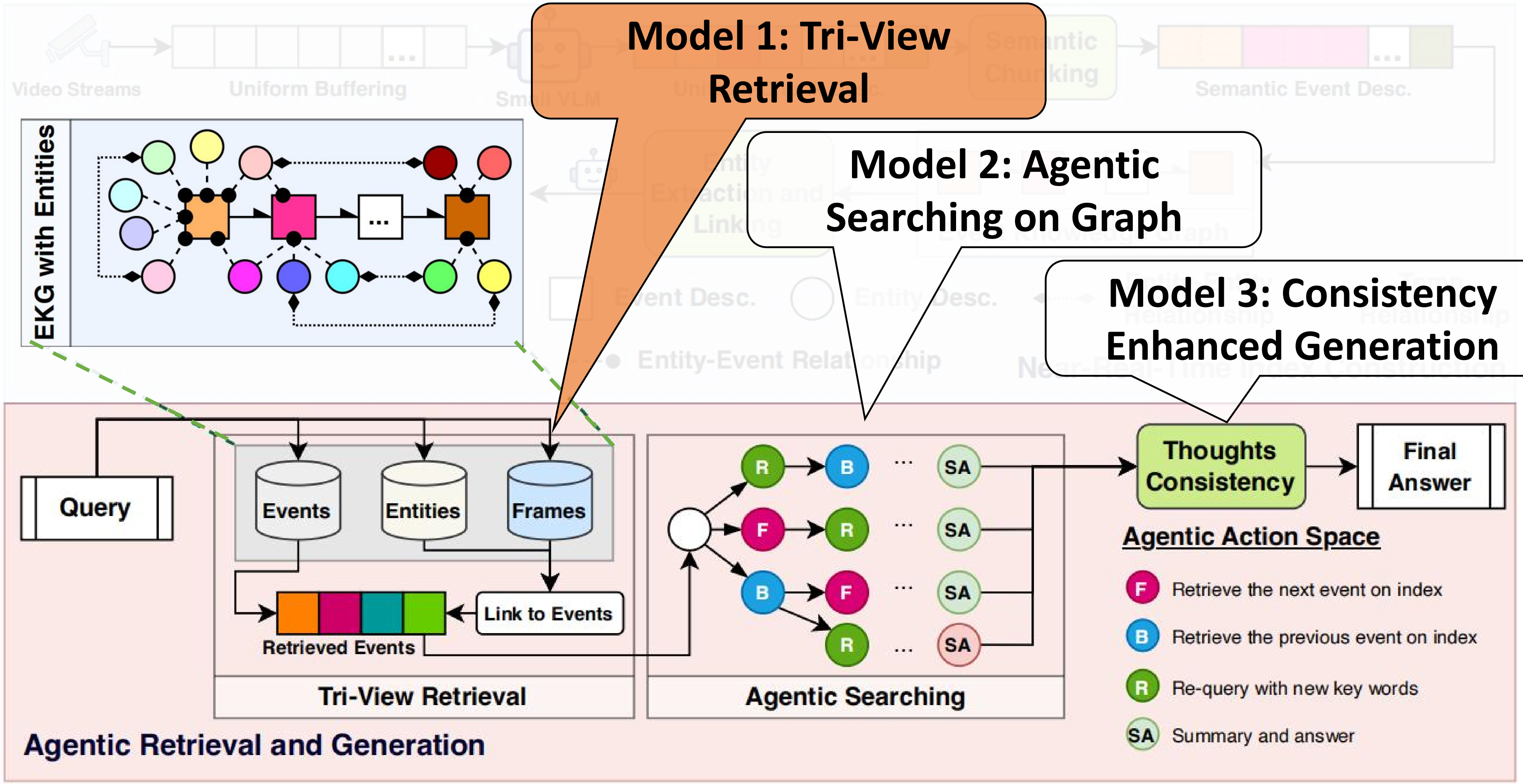
Three types of relationships:

1) semantic entity-entity relations

2) participation relations

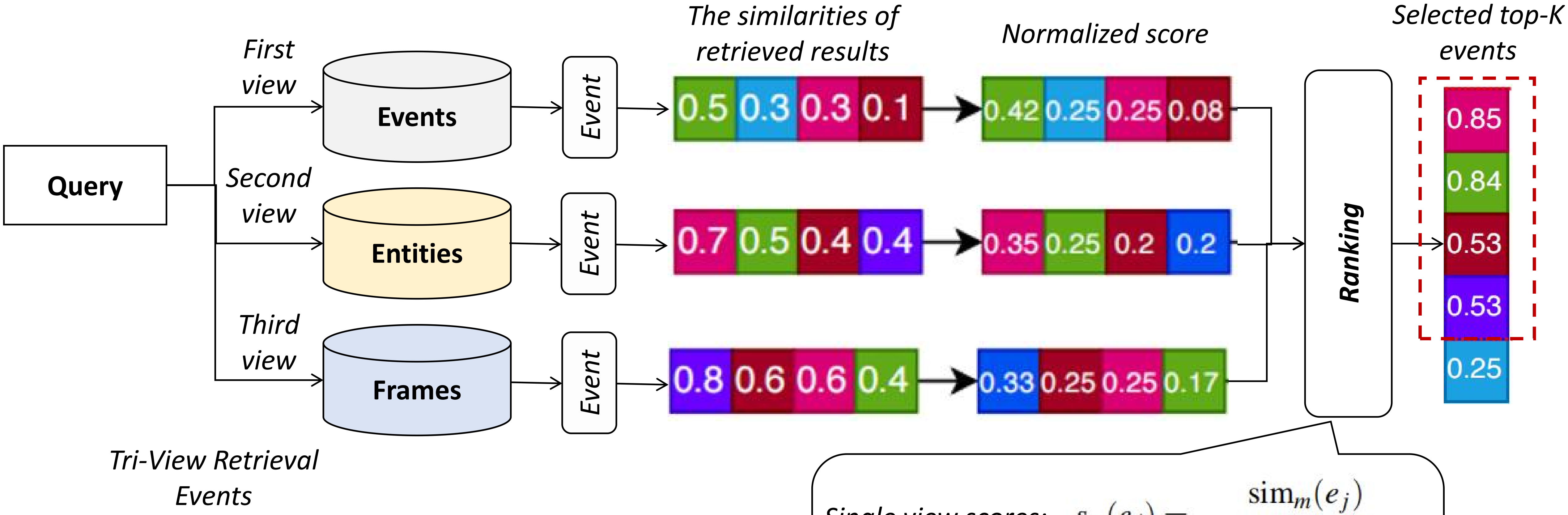
3) temporal event-event relations

Challenge 2: How to achieve accurate and efficient agentic searching on graph?



Model 1: Tri-View Retrieval

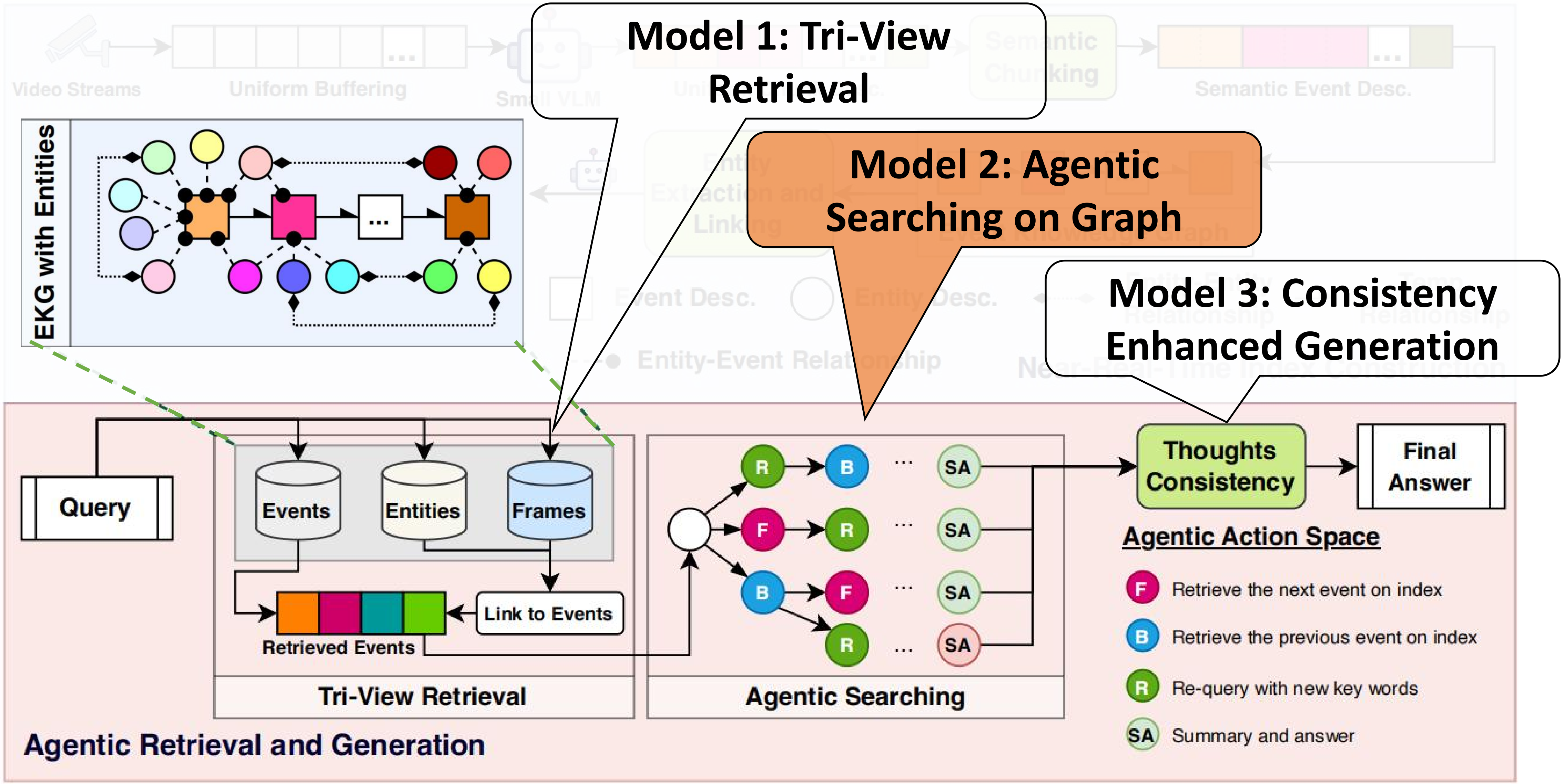
Key idea: *For a given query, simultaneous retrieval from three different views (i.e., event/entity/frame view) is performed to obtain more comprehensive and relevant information.*



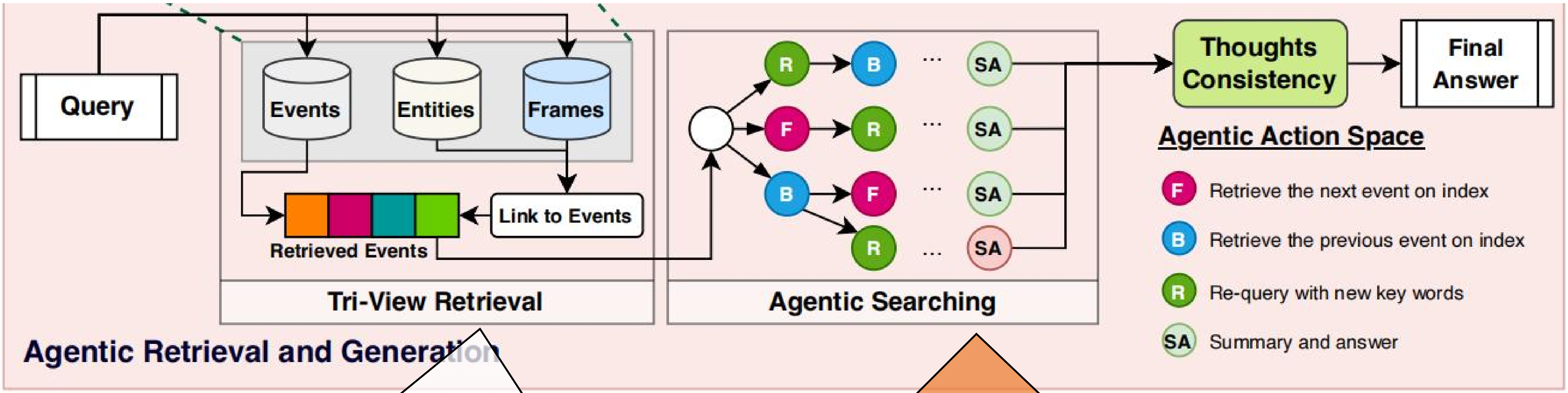
Single view scores: $$s_m(e_j) = \frac{\mathrm{sim}_m(e_j)}{\sum_{e_k \in \mathcal{E}_m} \mathrm{sim}_m(e_k)},$$

Tri-view scores: $$s(e_j) = \sum_m s_m(e_j),$$

Design 2: Agentic Retrieval and Generation



Model 2: Agentic Searching on Graph



Model 1: Tri-View Retrieval

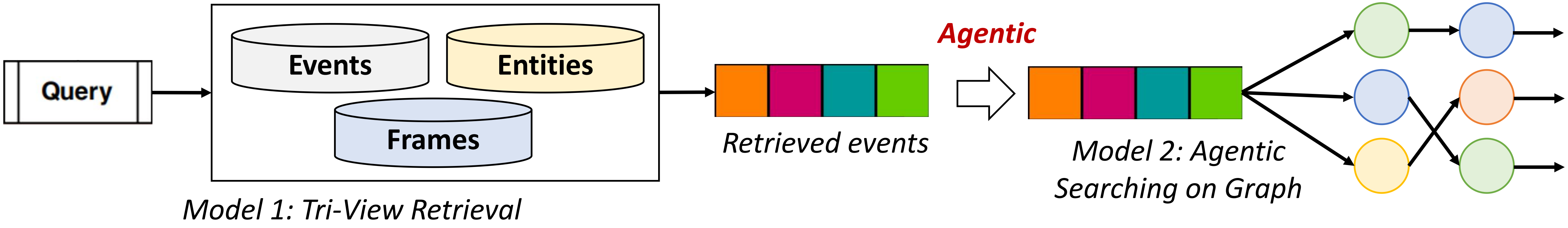
Agentic

Model 2: Agentic Searching on Graph

Closed-end query:
"Find all events with a raccoon"

Open-end query:
- *"Summarize all the abnormal activities that occurred in the past 10 hours."*
- *"What did the man do after he opened the fridge?"*

Model 2: Agentic Searching on Graph



Agentic action space:

F

1. Forward (F)

B

2. Backward (B)

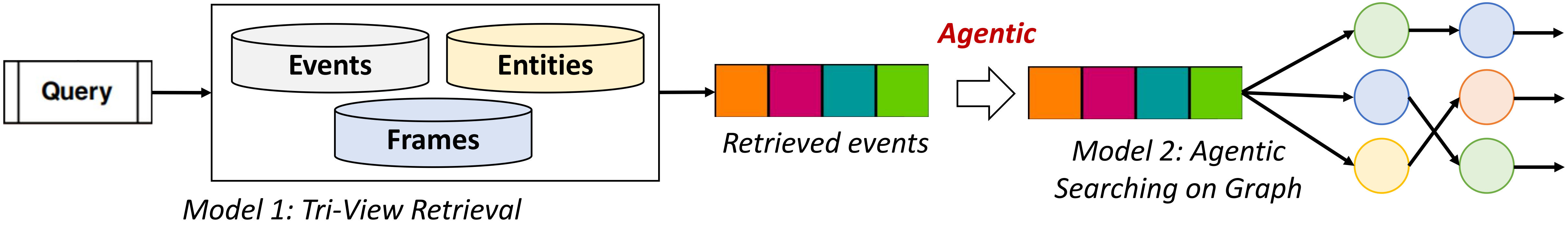
RQ

3. Re-query (RQ)

SA

4. Summary and Answer (SA)

Model 2: Agentic Searching on Graph

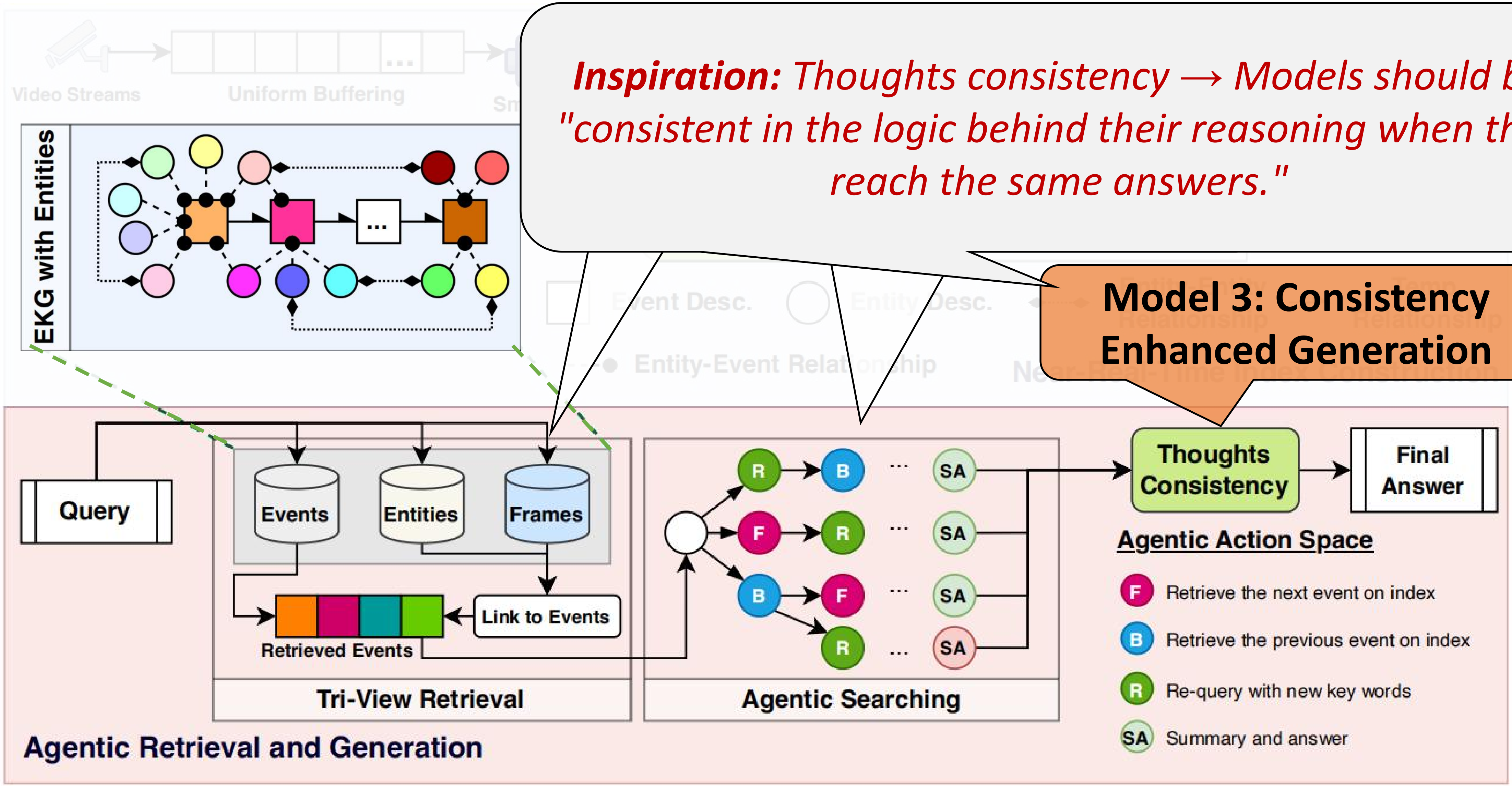


Qurey	Retrieved event	Agentic action space:	Enhanced event&describe
“那头独狼在发现废弃营地后，接下来去了哪里？”	”独狼-在营地边缘嗅探(时间: Day2 10:05)”	F 1. Forward (F)	“独狼-沿小溪向北移动(时间: Day2 10:28)”
“是什么原因导致鹿群在正午时分突然惊逃？”	“鹿群-突然惊逃-向东方(时间: Day3 12:15)”	B 2. Backward (B)	“金雕-从高空俯冲-掠过鹿群区域(时间: Day3 12:14)”
“评估一下该区域的人类活动强度。”	“研究人员-放置相机(时间: Day1 09:00)”	RQ 3. Re-query (RQ)	“评估一下该区域的人类活动强度，通过车、垃圾、灯光…”
“对比一下群体A和群体B的熊在觅食行为上的差异。”	“熊A-在树下翻找浆果；熊B-挖掘树根”	SA 4. Summary and Answer (SA)	“群体A更倾向于在白天…；群体B的觅食策略更多样化…”

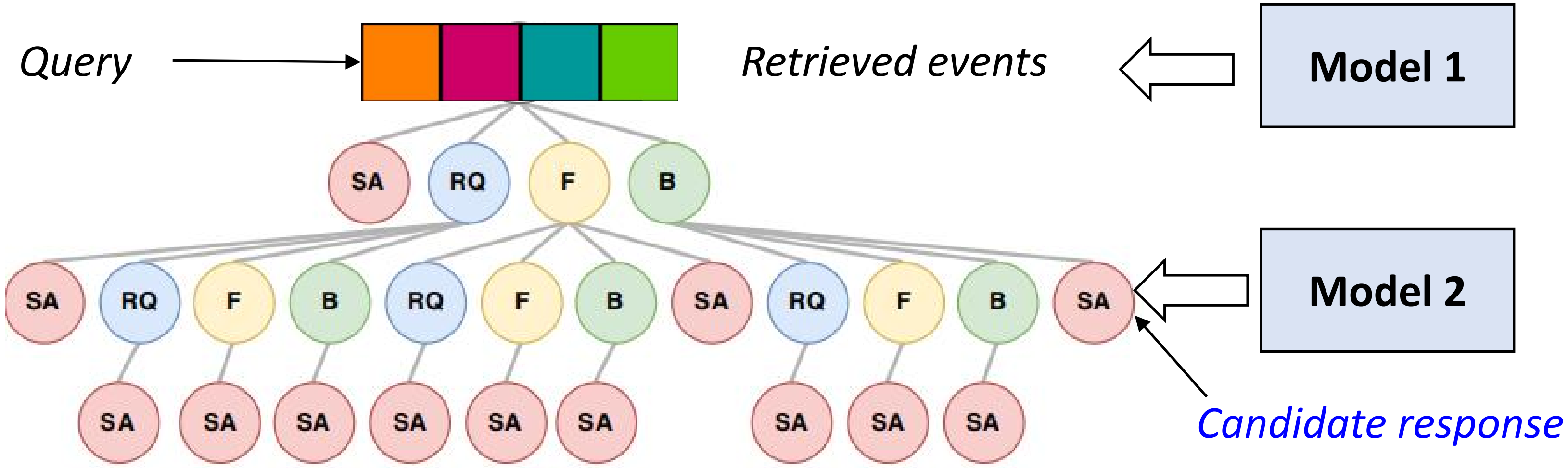
Overview

Inspiration: *Thoughts consistency → Models should be "consistent in the logic behind their reasoning when they reach the same answers."*

Model 3: Consistency Enhanced Generation



Model 3: Consistency Enhanced Generation



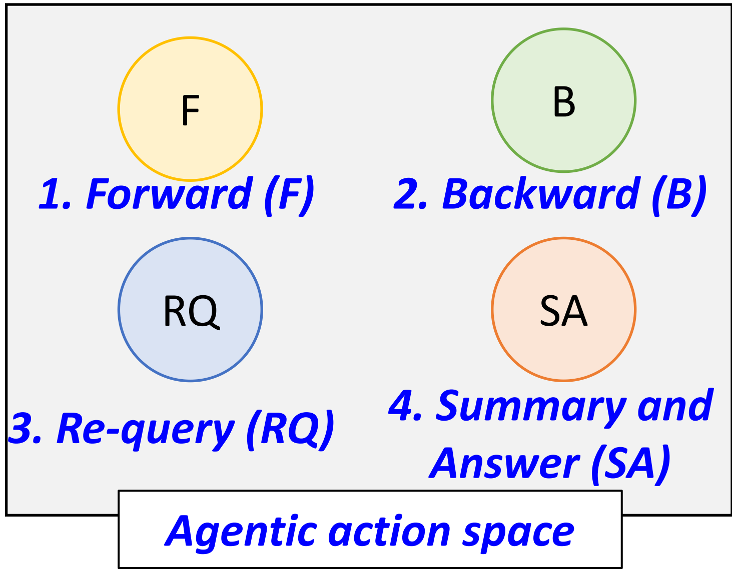
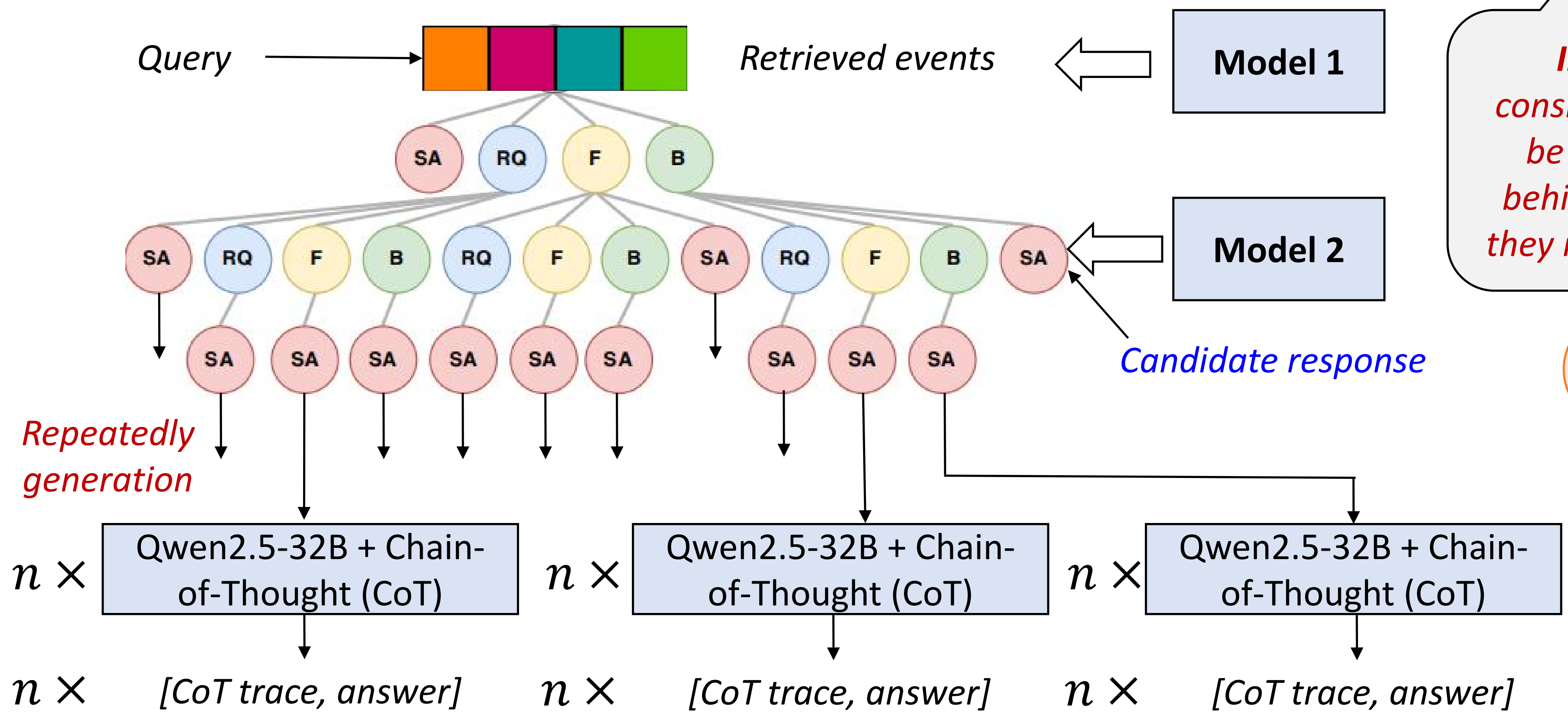


Figure 6: An example of agentic tree search with four actions and a depth of three, yielding 13 distinct pathways for information gathering and response generation.

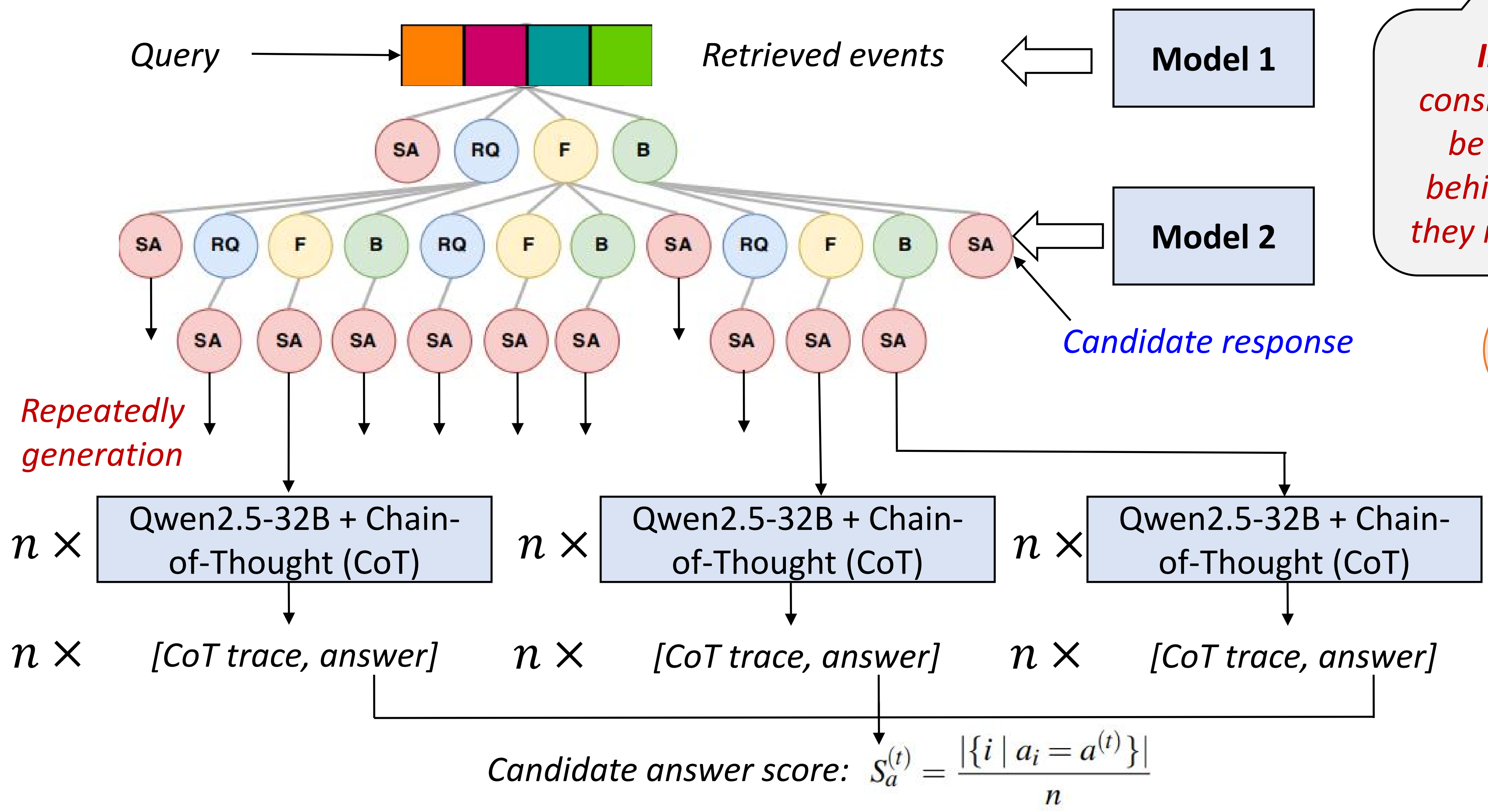
Model 3: Consistency Enhanced Generation



Inspiration: *Thoughts consistency → Models should be "consistent in the logic behind their reasoning when they reach the same answers."*

SA ***4. Summary and Answer (SA)***

Model 3: Consistency Enhanced Generation

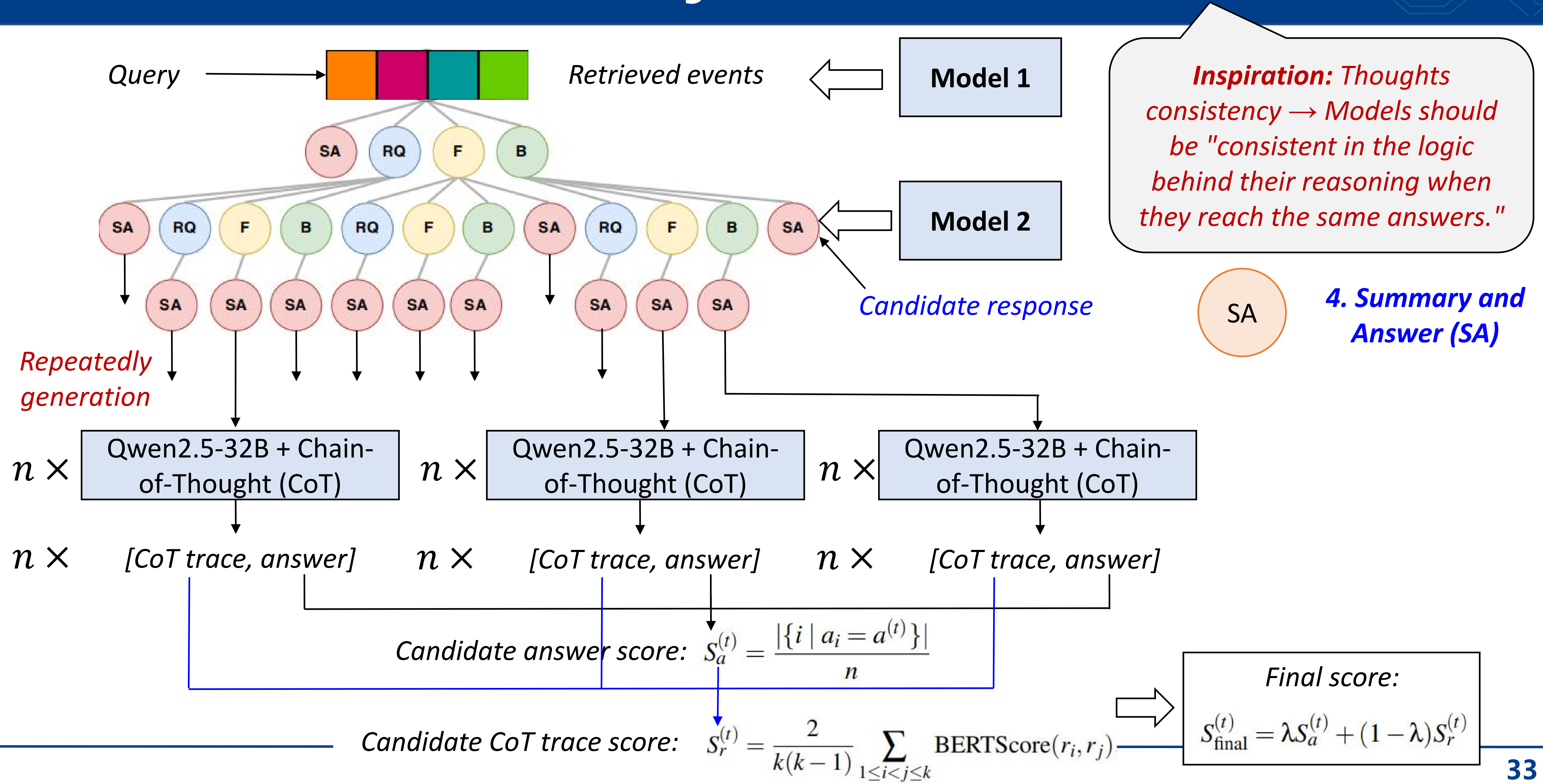


Inspiration: *Thoughts consistency → Models should be "consistent in the logic behind their reasoning when they reach the same answers."*

SA

4. Summary and Answer (SA)

Model 3: Consistency Enhanced Generation



Outline

1 Background and SOTA
2 Opportunity and Challenge
3 Design
4 Evaluation
5 Conclusion

Evaluation Settings

Benchmarks:

- LVBench [ICCV’25]
  - 103 videos (4100 seconds/video)
  - 1549 questions
- VideoMME-Long [CVPR’25]
  - 300 videos (2400 seconds/video)
  - 900 questions
- AVA-100 [NSDI'26]
  - 8 videos (10 hours/video)
  - 120 questions

Baselines:

- VLM
  - GPT-4o
  - Gemini-1.5-Pro
  - Phi- 4-Multimodal
  - Qwen2.5-VL-7B
  - InternVL2.5-8B
  - LLaVA-Video-7B
- Video-RAG method
  - VideoTree [CVPR’25]
  - VideoAgent [ECCV’24]
  - DrVideo [CVPR’25]
  - VCA [ICCV’25]

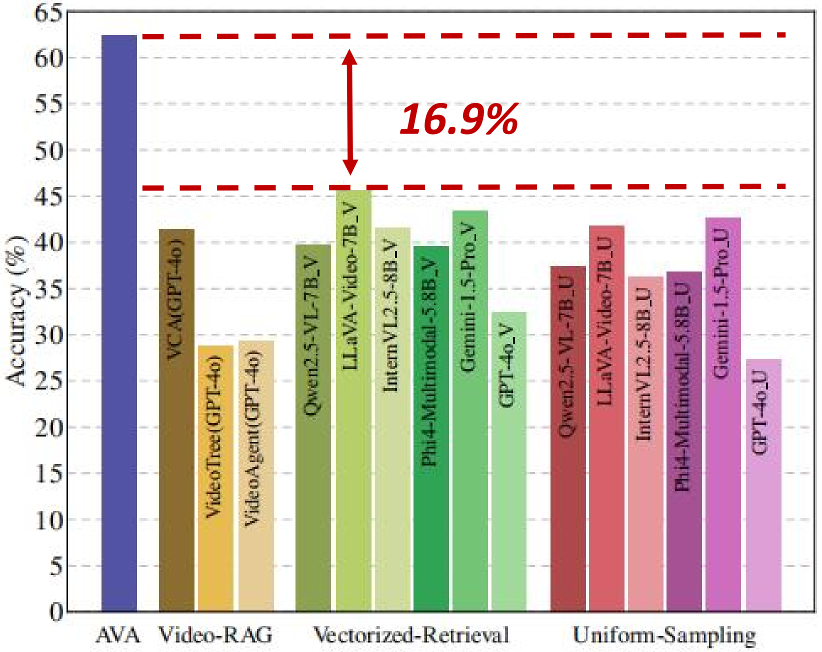
×

Two typical strategies:

- Uniform sampling
- Vectorized retrieval (top-K)

[1] [ICCV’25] Lvbench: An extreme long video understanding benchmark. In International Conference on Computer Vision
[2] [CVPR’25] Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis
[3] [NSDI'26] AVA: Towards Agentic Video Analytics with Vision Language Models
[4] [CVPR’25] Videotree: Adaptive tree based video representation for llm reasoning on long videos
[5] [ECCV’24] Videoagent: Long-form video understanding with large language model as agent
[6] [CVPR’25] Drvideo: Document retrieval based long video understanding
[7] [ICCV’25] Vca: Video curious agent for long video understanding

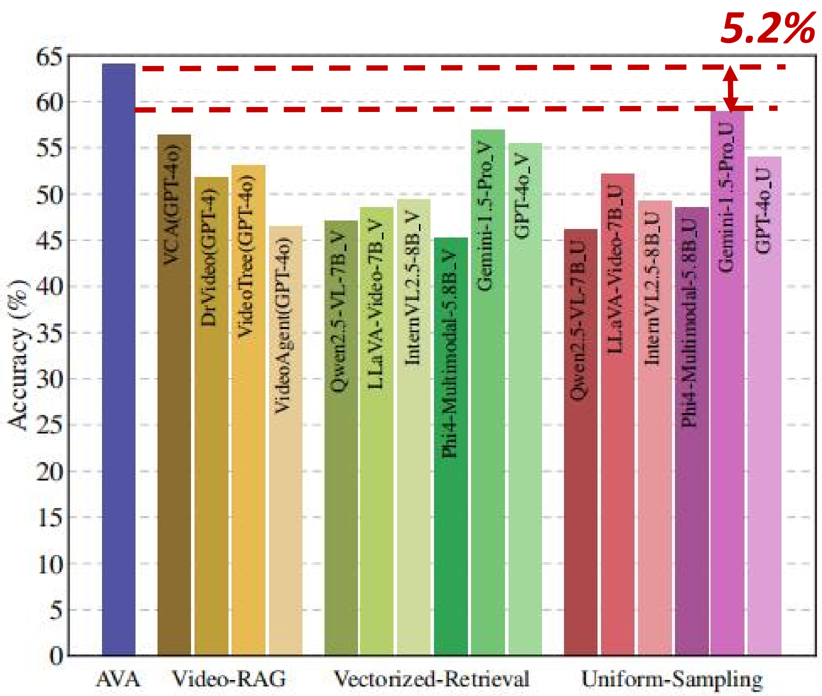
Overall Evaluation



Baseline 1: Video-RAG

Baseline 2: VLM

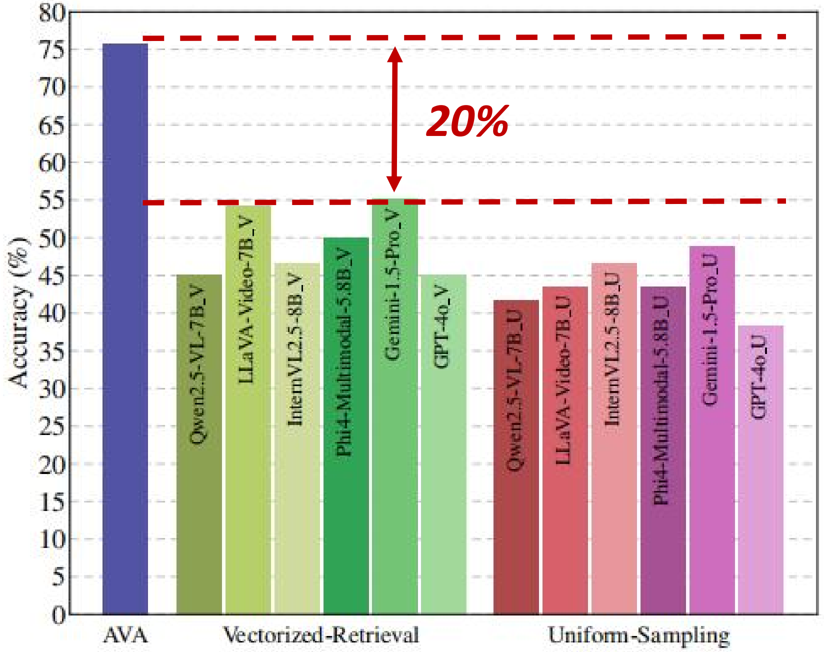
Benchmark 1: LVBench



Baseline 1: Video-RAG

Baseline 2: VLM

Benchmark 2: VideoMME-Long

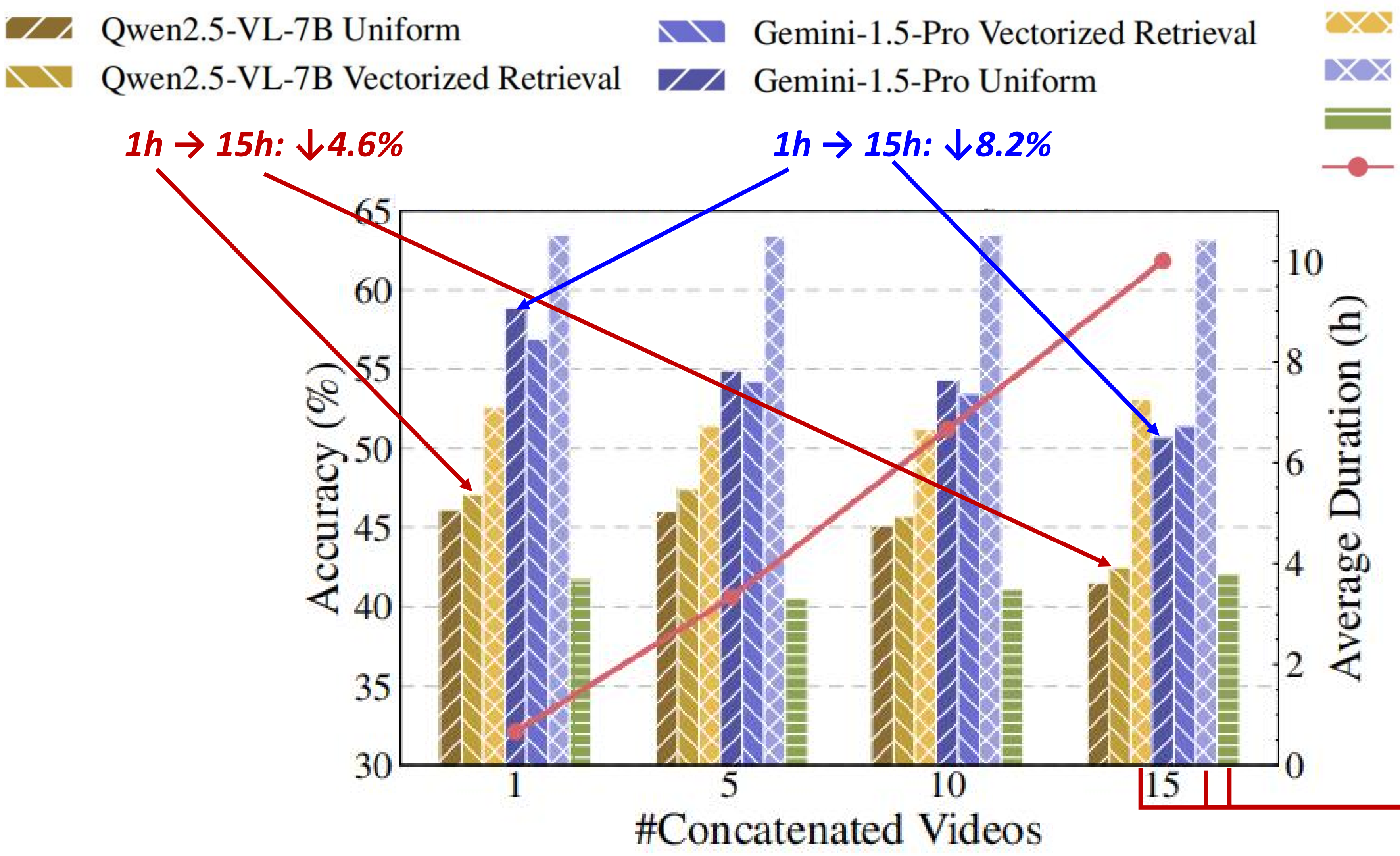


Baseline 2: VLM

Benchmark 3: AVA-100

AVA maintains robust performance for handling L4 video analytics tasks

Performance on Different Video Lengths



AVA consistently maintains stable performance across all video lengths

Figure: The accuracy achieved by AVA and the baselines across varying video lengths via concatenating videos from LVBench.

Outline

1 Background and SOTA

2 Opportunity and Challenge

3 Design

4 Evaluation

5 Conclusion

Conclusion

- AVA, the first L4 video analytics system powered by VLMs, to the best of our knowledge.
- Near-real-time index construction and agentic retrieval and generation, enabling open-end analytics on extremely long video sources in near-real-time.
- AVA-100, a benchmark specifically designed for L4 video analytics systems

Thinking

- **Apply to our scenario**
  - Event Knowledge Graph
  —> Hierarchical memory (Episodic/semantic memory)
  —> Graph continual learning
  - Agentic Searching (action: F, B, RQ, SA)
  —> APIs: Resource allocation parameter search
- **More general**
  - [NSDI'26] AVA —> edgeAVA
  - Event Knowledge Graph (video/language) —> Multimodal data
- **Direct improvement**
  - F, B, RQ, SA —> More agentic action space
  - Near-real-time index construction (>2FPS) —> More real time

Thanks for your learning
Q & A