Q_2E: **Q**uery-to-**E**vent Decomposition for Zero-Shot Multilingual Text-to-Video Retrieval

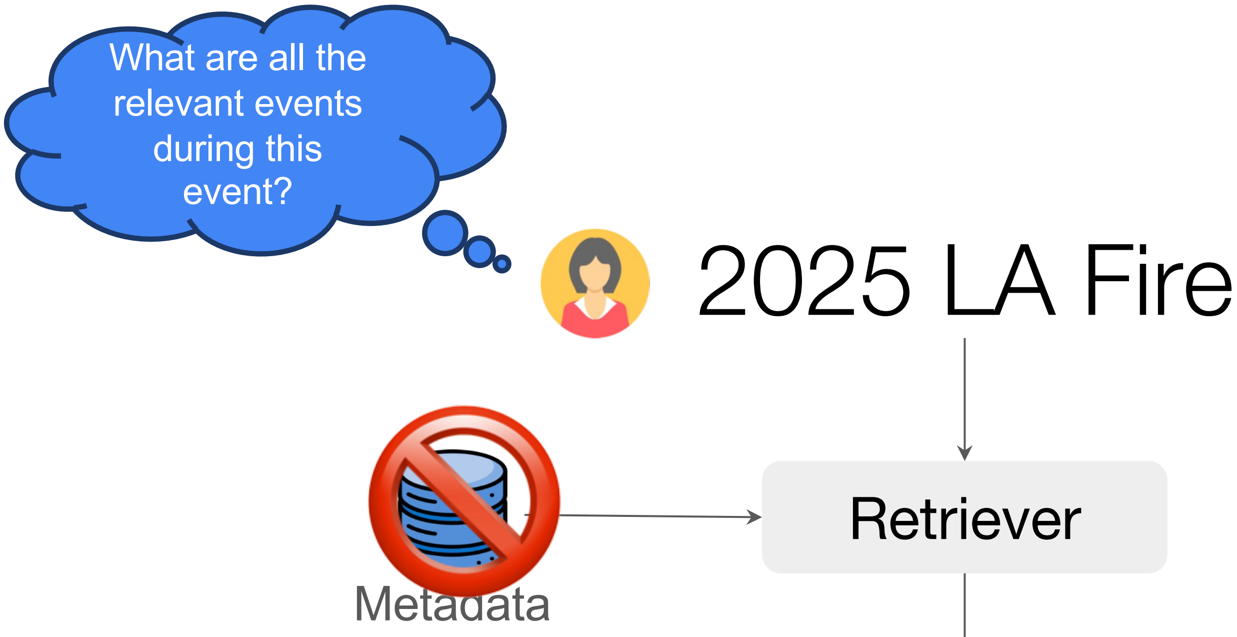
Shubhashis Roy Dipta

Frank Ferraro

UMBC

Retrieving videos is hard **without** metadata. How can we improve it…

- via enriching the query?
- via information *within* the video?



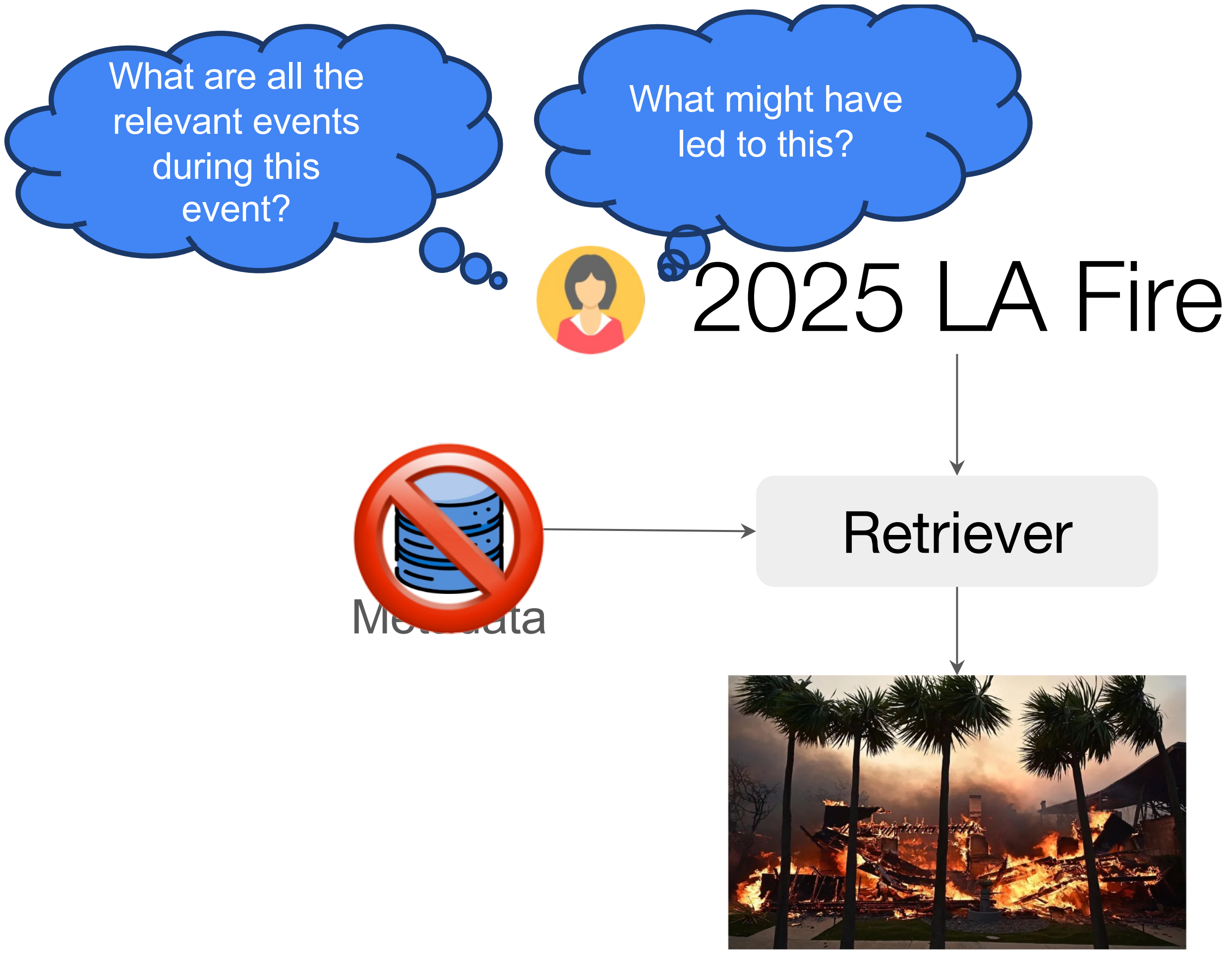
What are all the relevant events during this event?

2025 LA Fire

Retriever

Metadata

Originally: 2025 LA Fire



What are all the relevant events during this event?

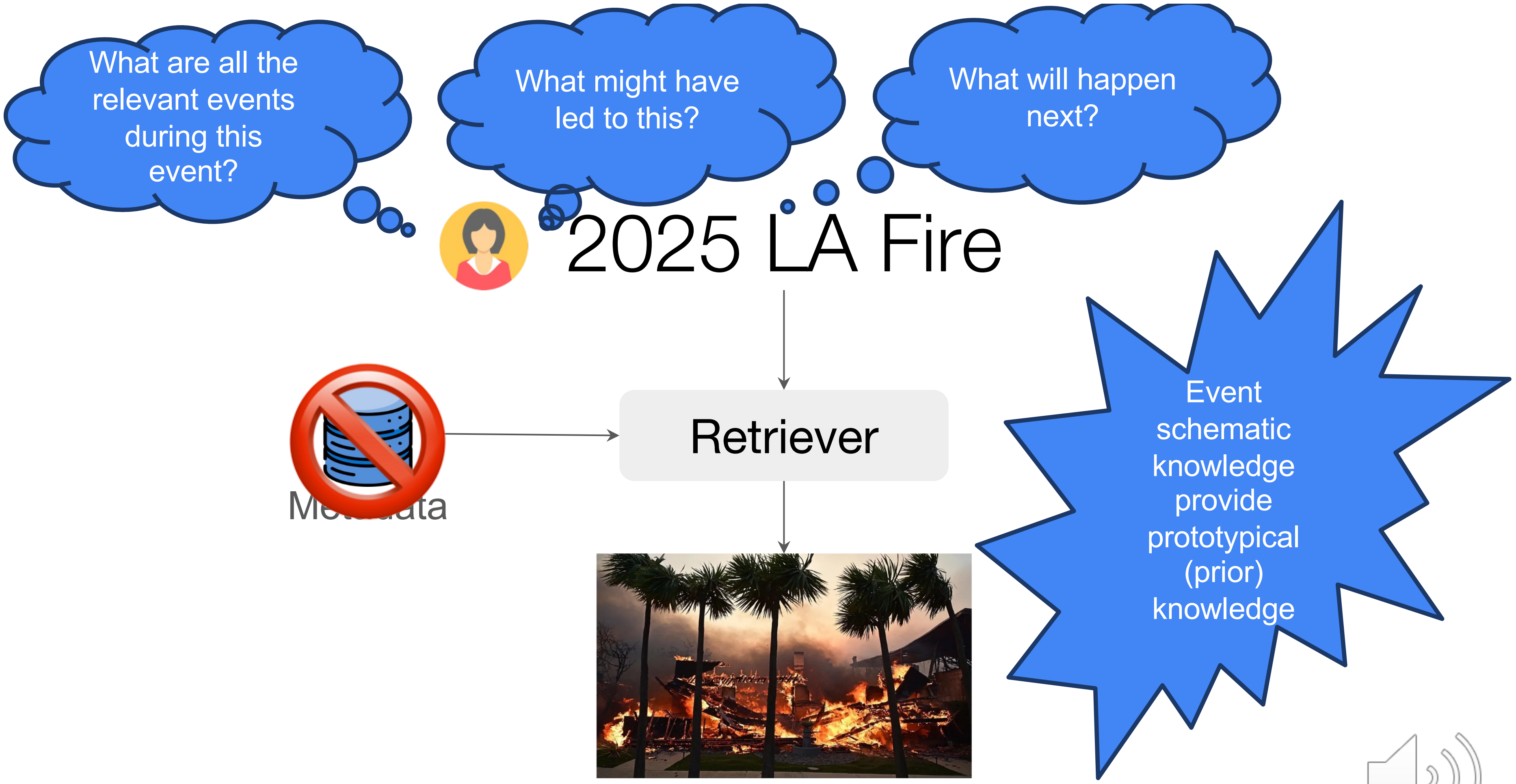
What might have led to this?

2025 LA Fire

Retriever

Metadata

Originally: 2025 LA Fire



What are all the relevant events during this event?

What might have led to this?

What will happen next?

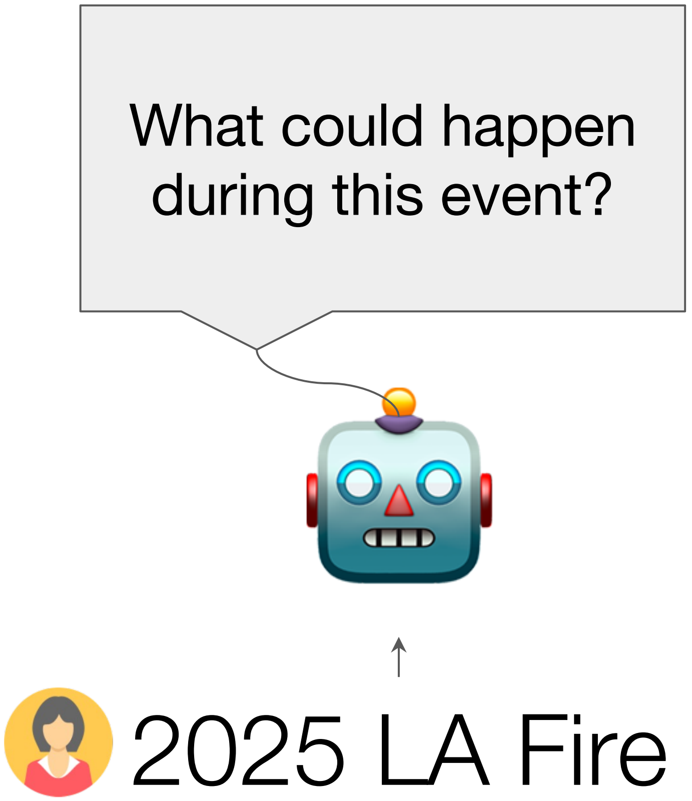
2025 LA Fire

Metadata

Retriever

Event schematic knowledge provide prototypical (prior) knowledge

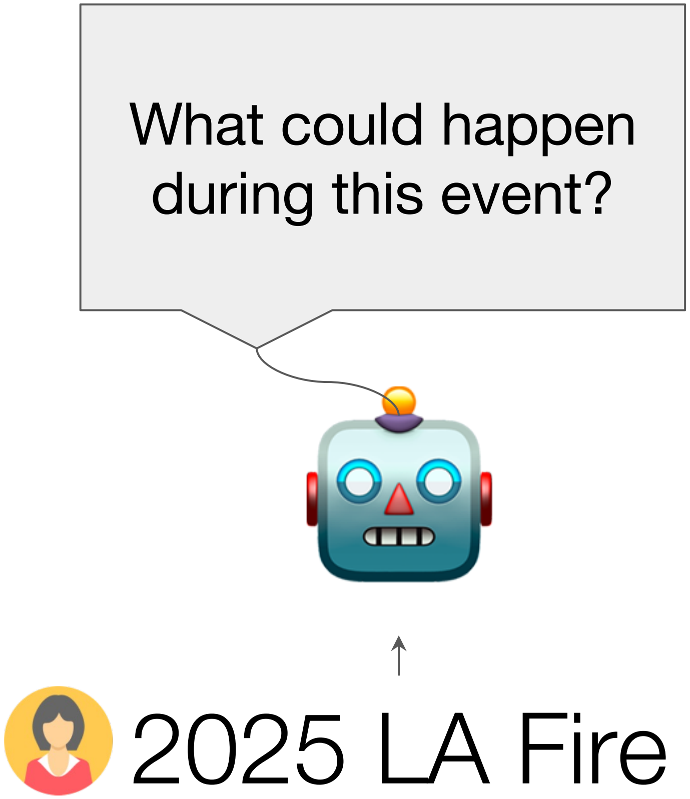
Originally: 2025 LA Fire



What could happen during this event?

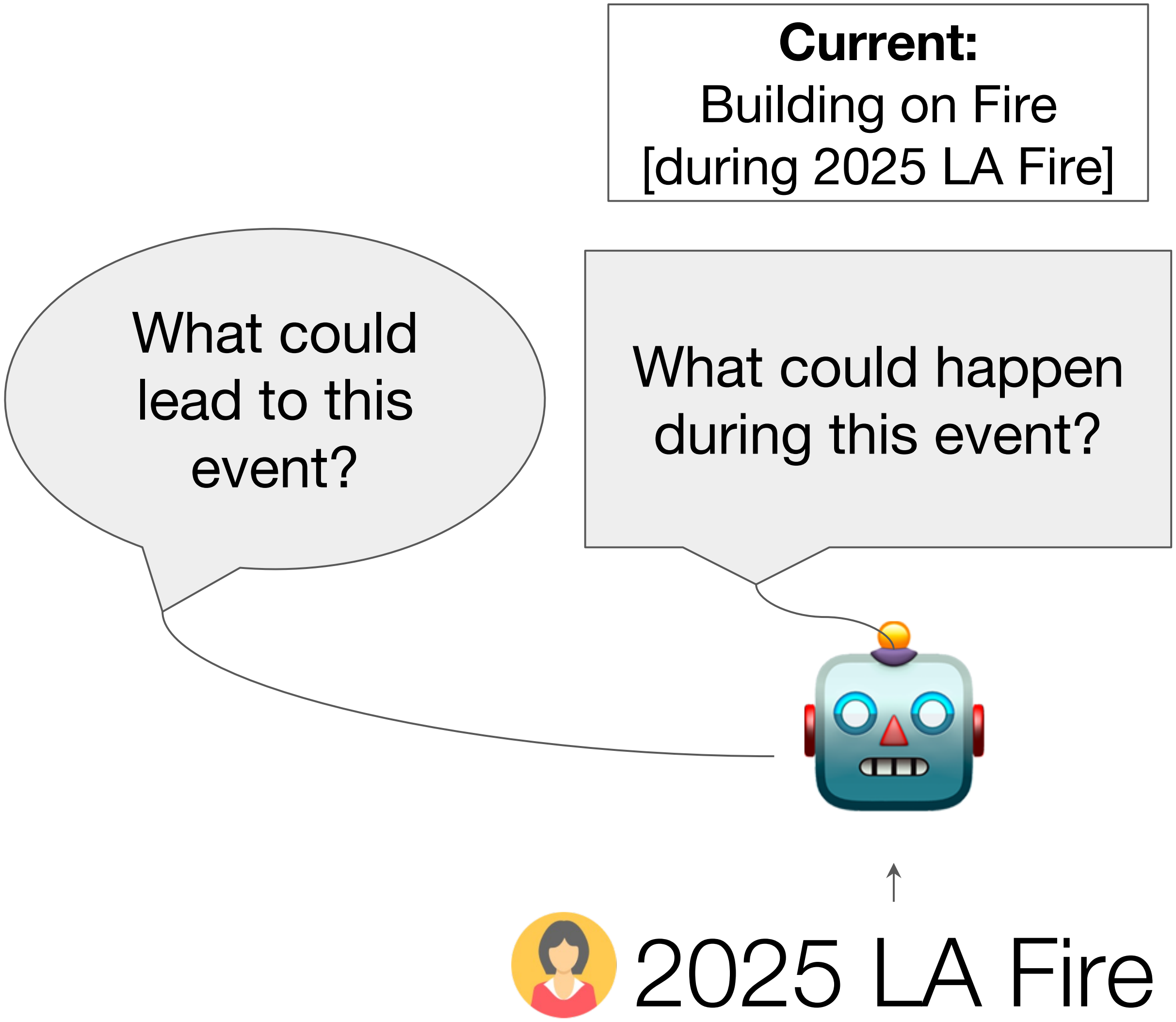
2025 LA Fire

Current:
Building on Fire
[during 2025 LA Fire]



What could happen during this event?

2025 LA Fire

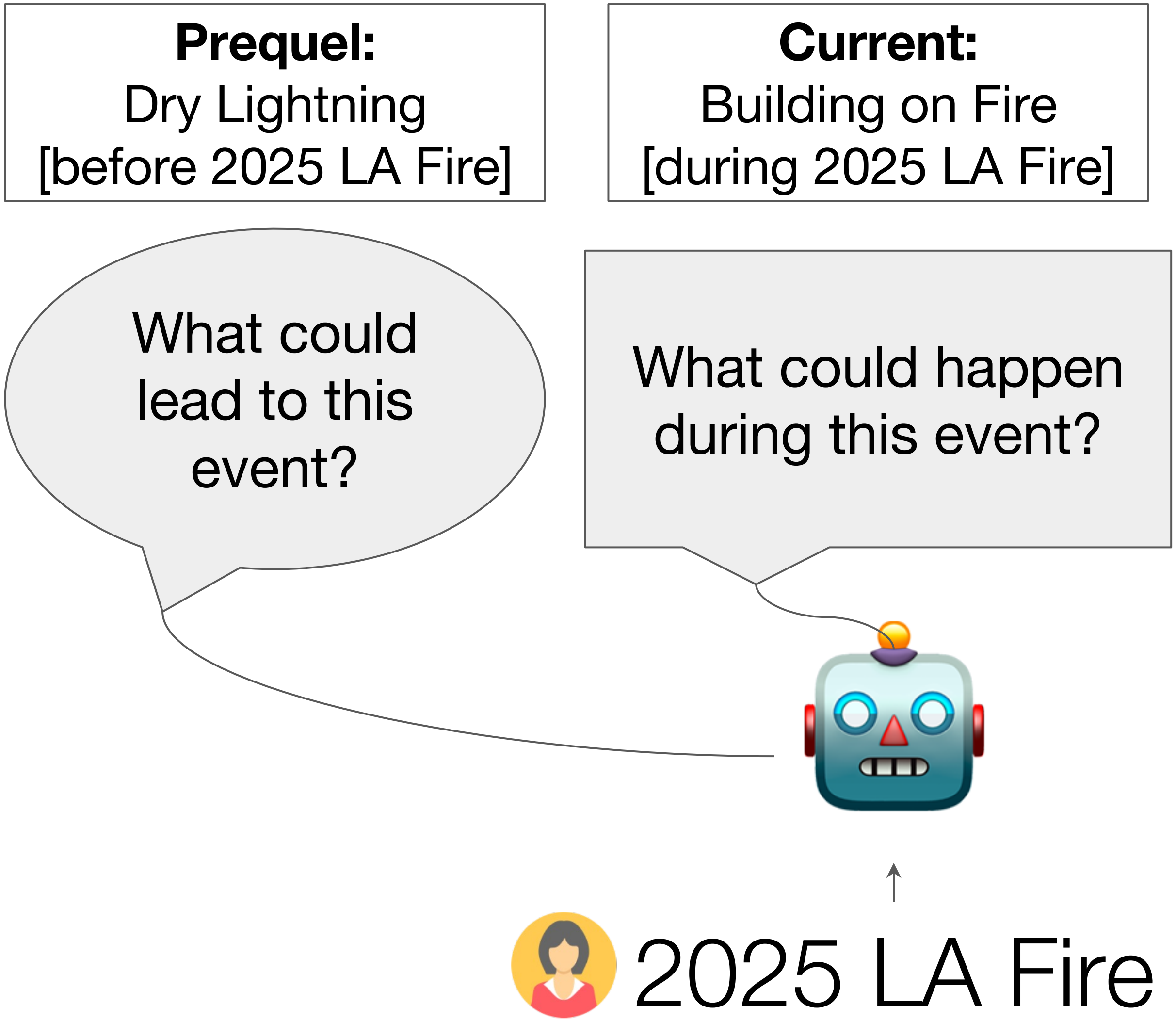


Current:
Building on Fire
[during 2025 LA Fire]

What could lead to this event?

What could happen during this event?

2025 LA Fire



Prequel:
Dry Lightning
[before 2025 LA Fire]

Current:
Building on Fire
[during 2025 LA Fire]

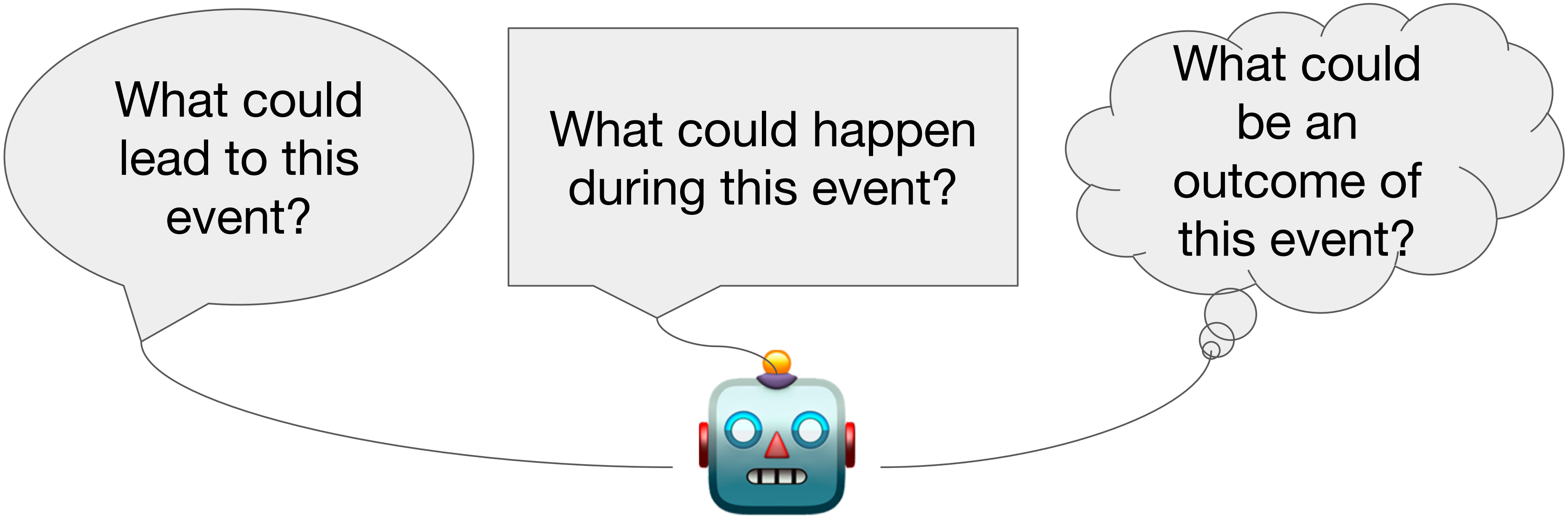
What could lead to this event?

What could happen during this event?

2025 LA Fire

Prequel:
Dry Lightning
[before 2025 LA Fire]

Current:
Building on Fire
[during 2025 LA Fire]



What could lead to this event?

What could happen during this event?

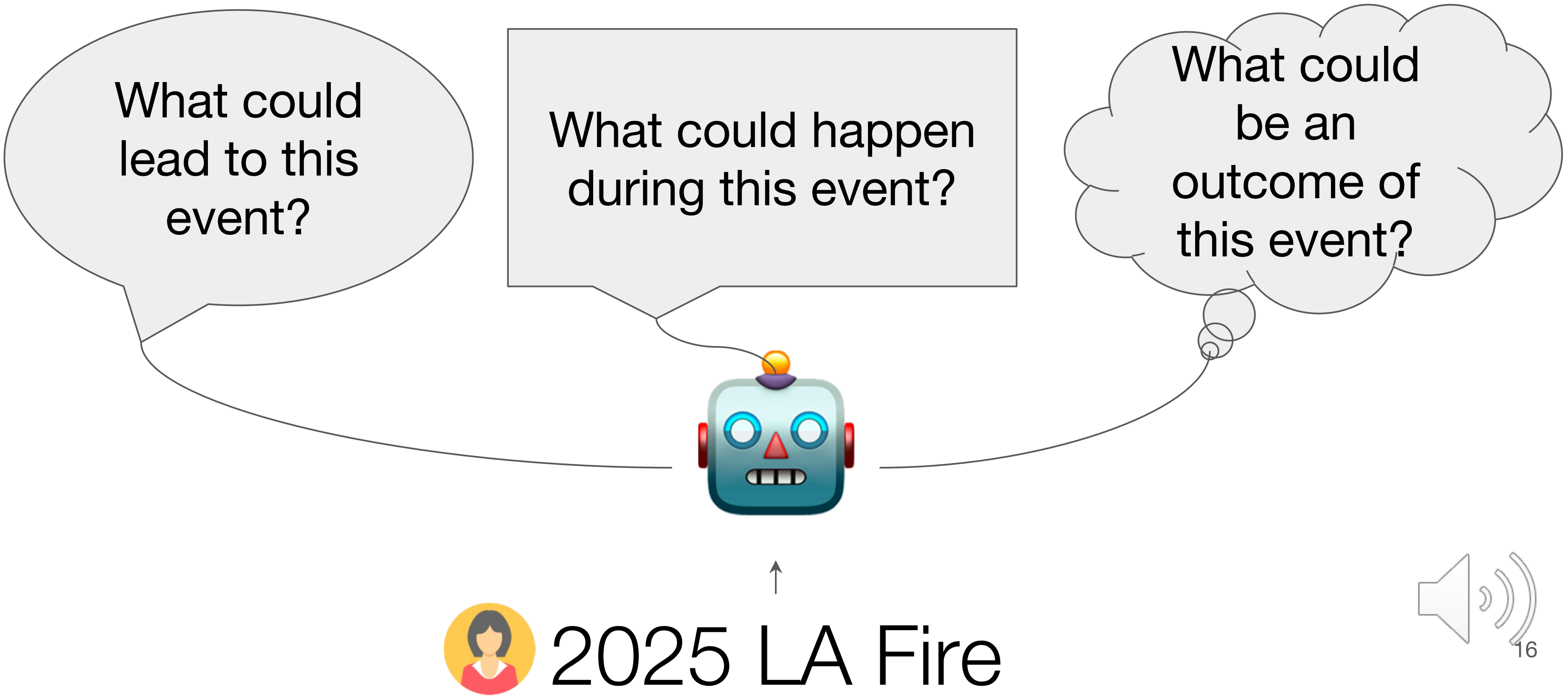
What could be an outcome of this event?

2025 LA Fire

Prequel:
Dry Lightning
[before 2025 LA Fire]

Current:
Building on Fire
[during 2025 LA Fire]

Sequel:
People are returning
[after 2025 LA Fire]

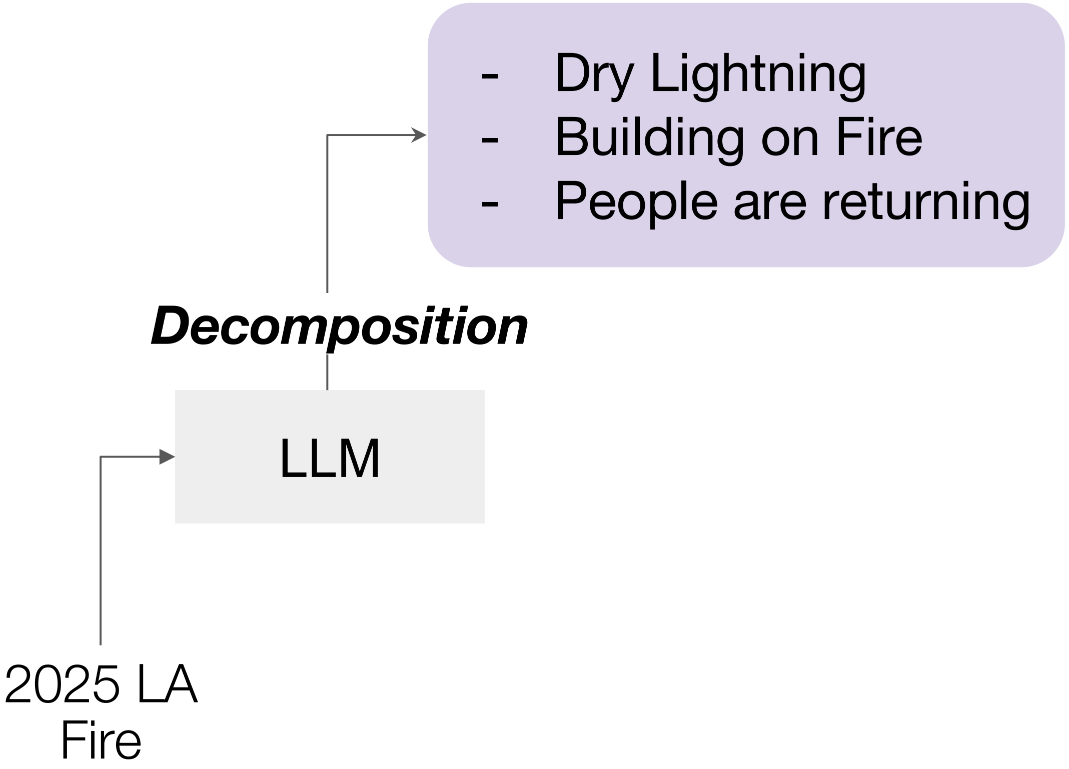


What could lead to this event?

What could happen during this event?

What could be an outcome of this event?

2025 LA Fire

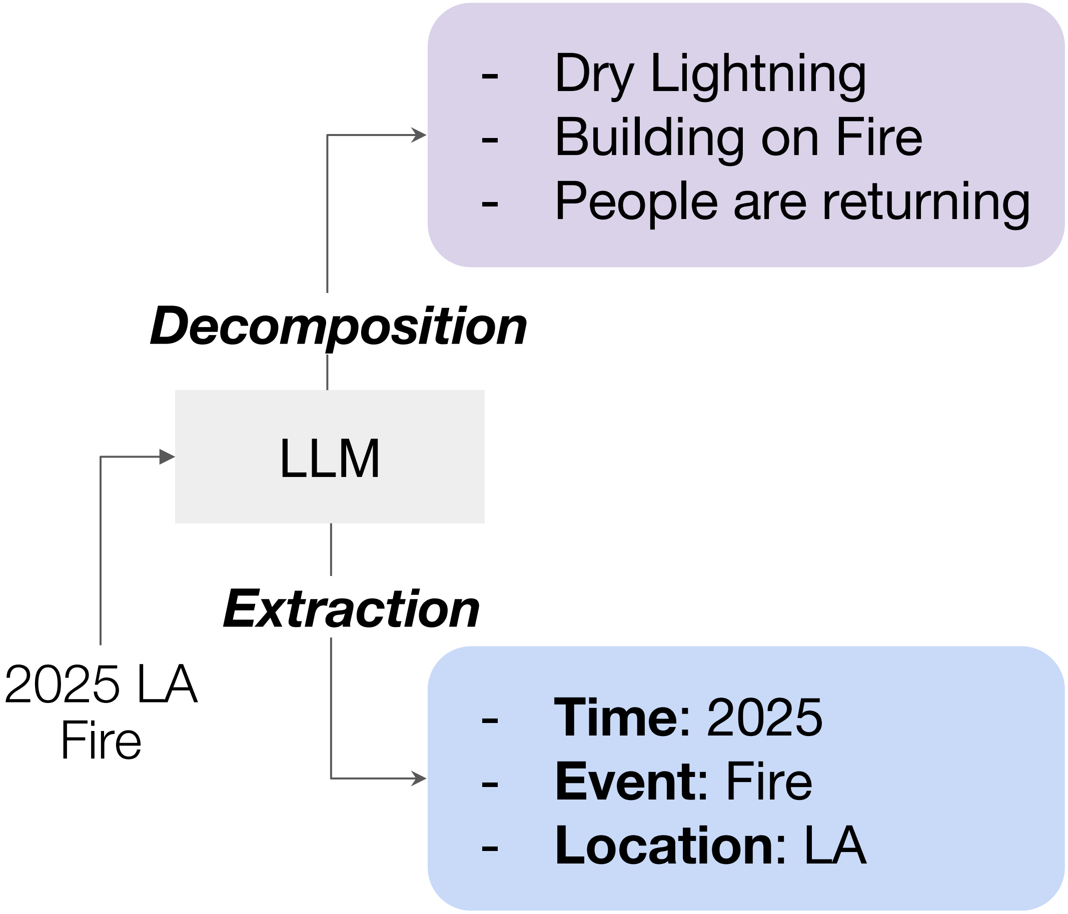


- Dry Lightning
- Building on Fire
- People are returning

Decomposition

LLM

2025 LA Fire



2025 LA Fire

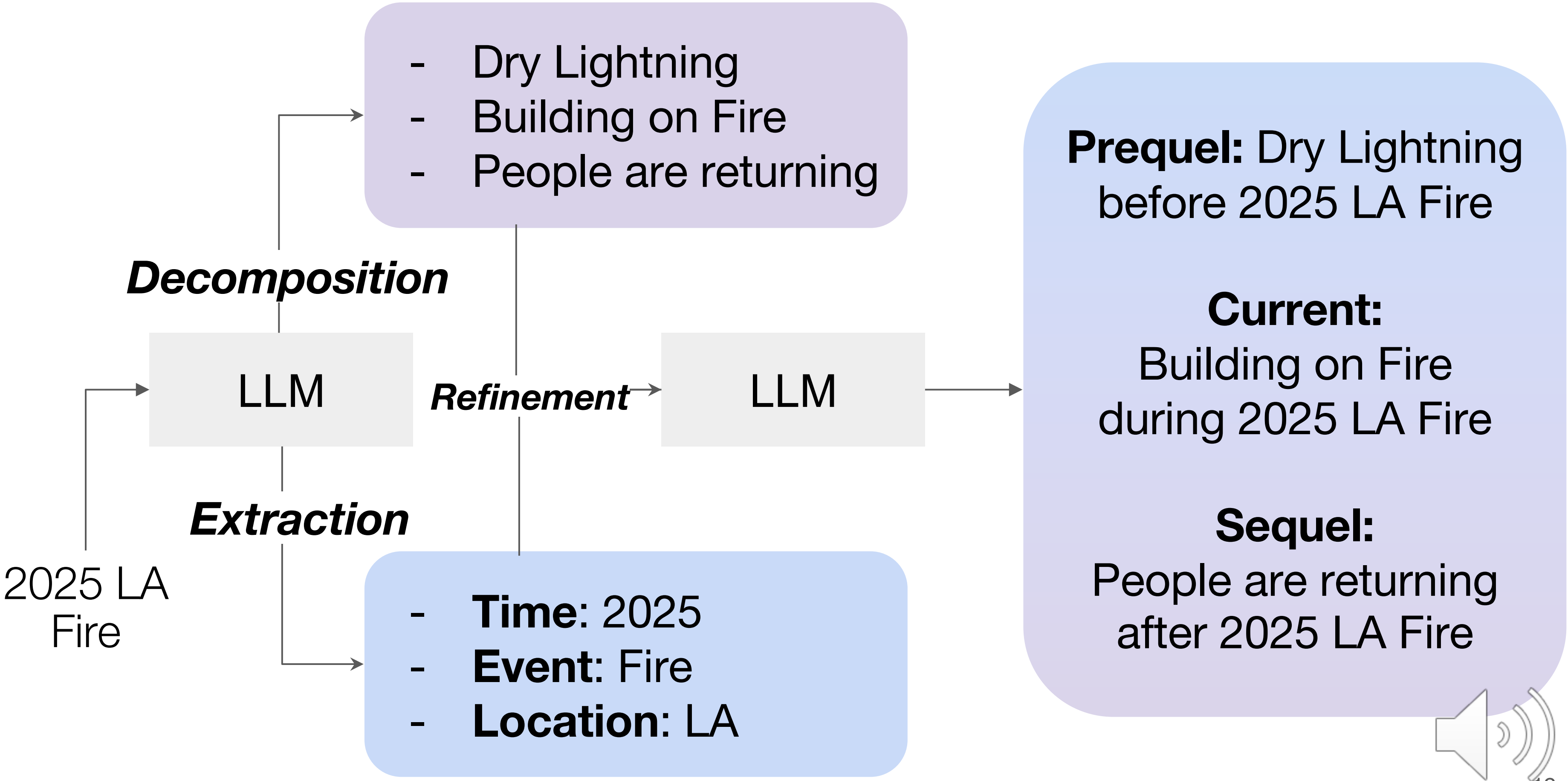
LLM

Decomposition

- Dry Lightning
- Building on Fire
- People are returning

Extraction

- **Time**: 2025
- **Event**: Fire
- **Location**: LA



2025 LA Fire

LLM

Decomposition

- Dry Lightning
- Building on Fire
- People are returning

Extraction

- **Time**: 2025
- **Event**: Fire
- **Location**: LA

Refinement

LLM

Prequel: Dry Lightning before 2025 LA Fire

Current: Building on Fire during 2025 LA Fire

Sequel: People are returning after 2025 LA Fire

Retrieving videos is hard **without** metadata. How can we improve it…

- via enriching the query? ✅
- via information *within* the video?

Query: 2025 LA Fire

Video Desc.:
Trees and a building are on fire. … →
A wildfire burning.

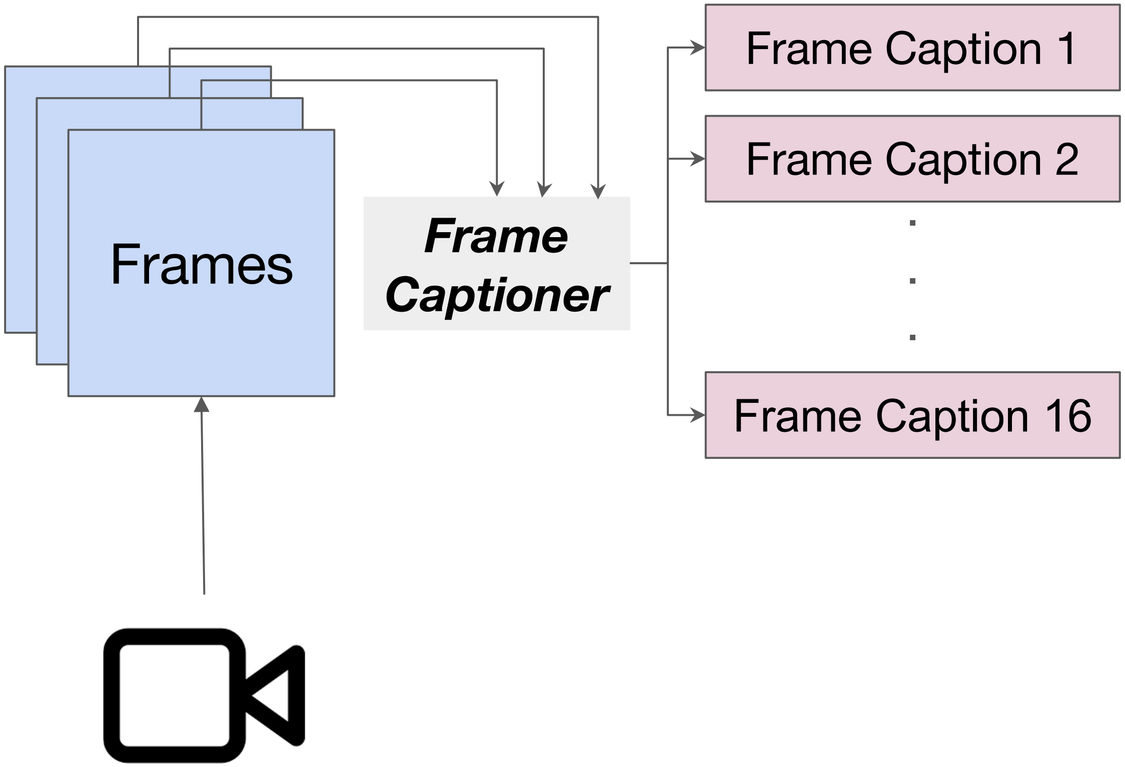
ASR:
Today, Aug 16, 2020
a massive wildfire
has started

Enrichment from Videos

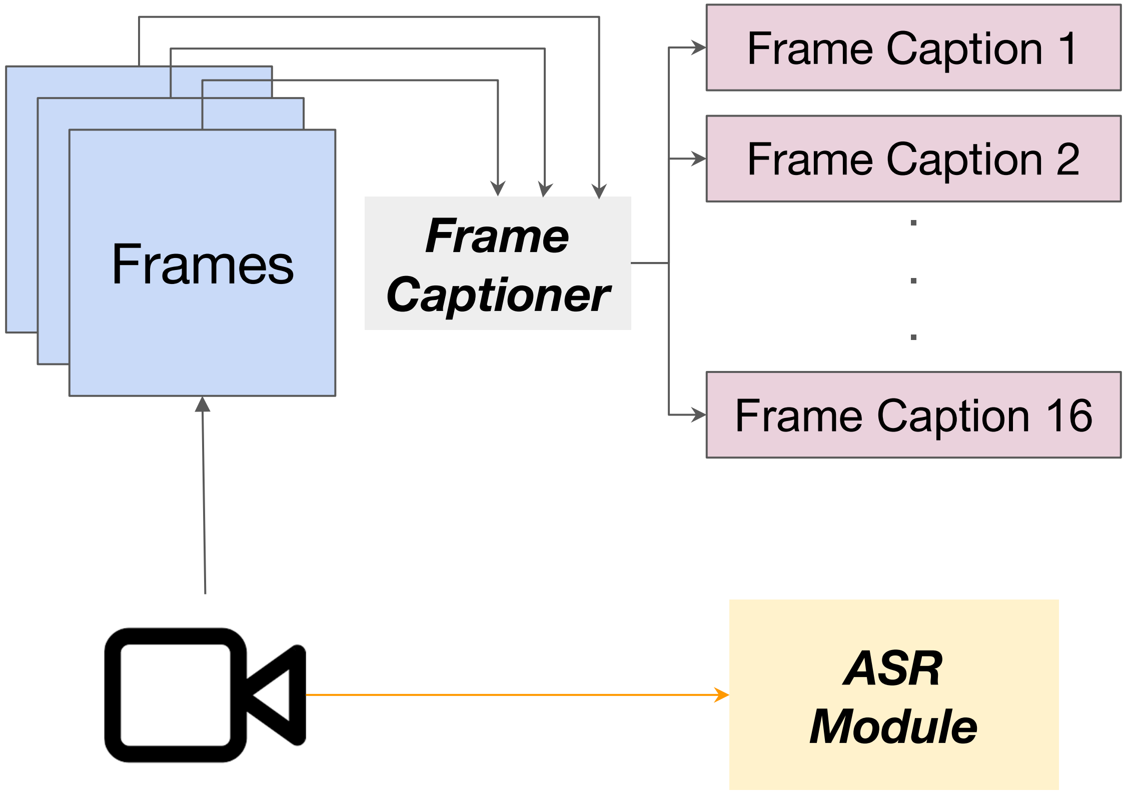
Can’t VLMs just do this?

- Descriptions may not be at the right level of granularity
- Current VLM fails to understand the full context
- Current ASR fails to both translate and transcribe in one go
- Multilingual videos still pose challenges

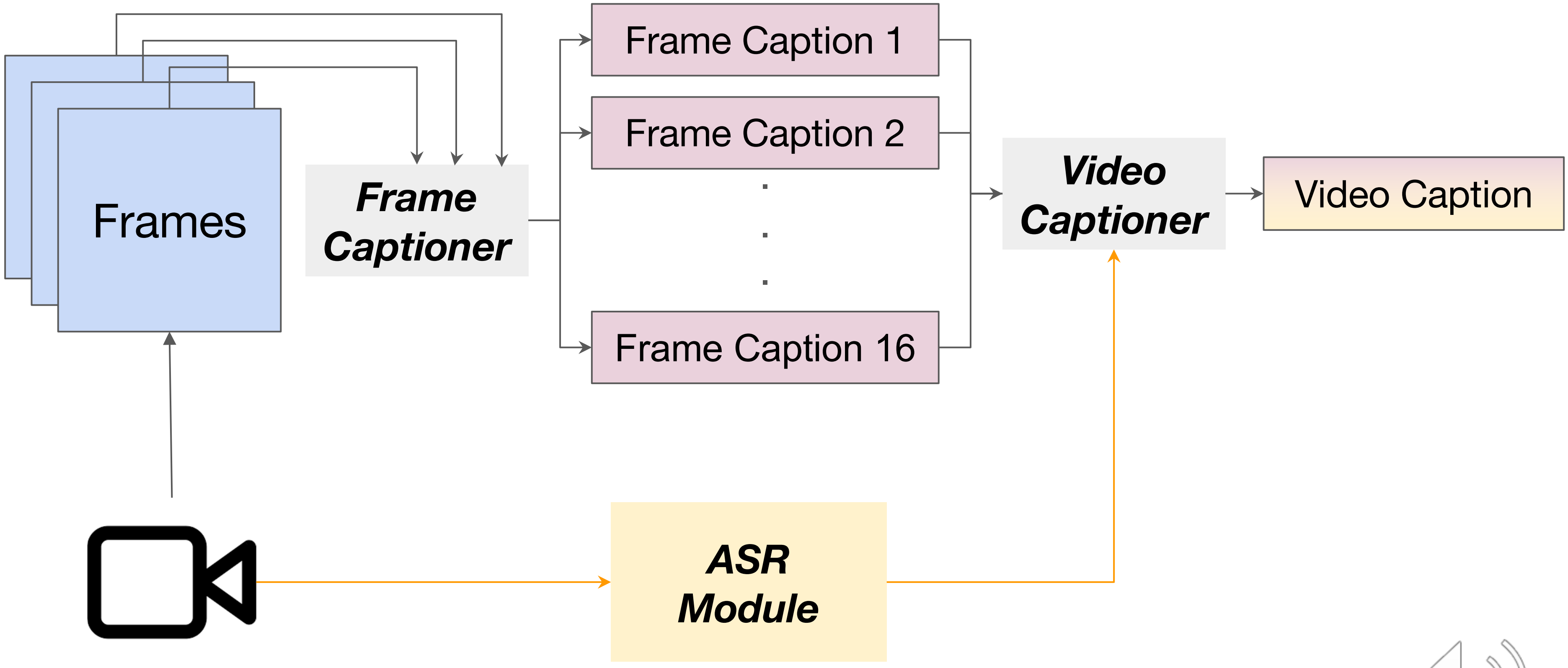
Video Description Extraction



Video Description Extraction



Video Description Extraction



Frames

Frame Captioner

Frame Caption 1

Frame Caption 2

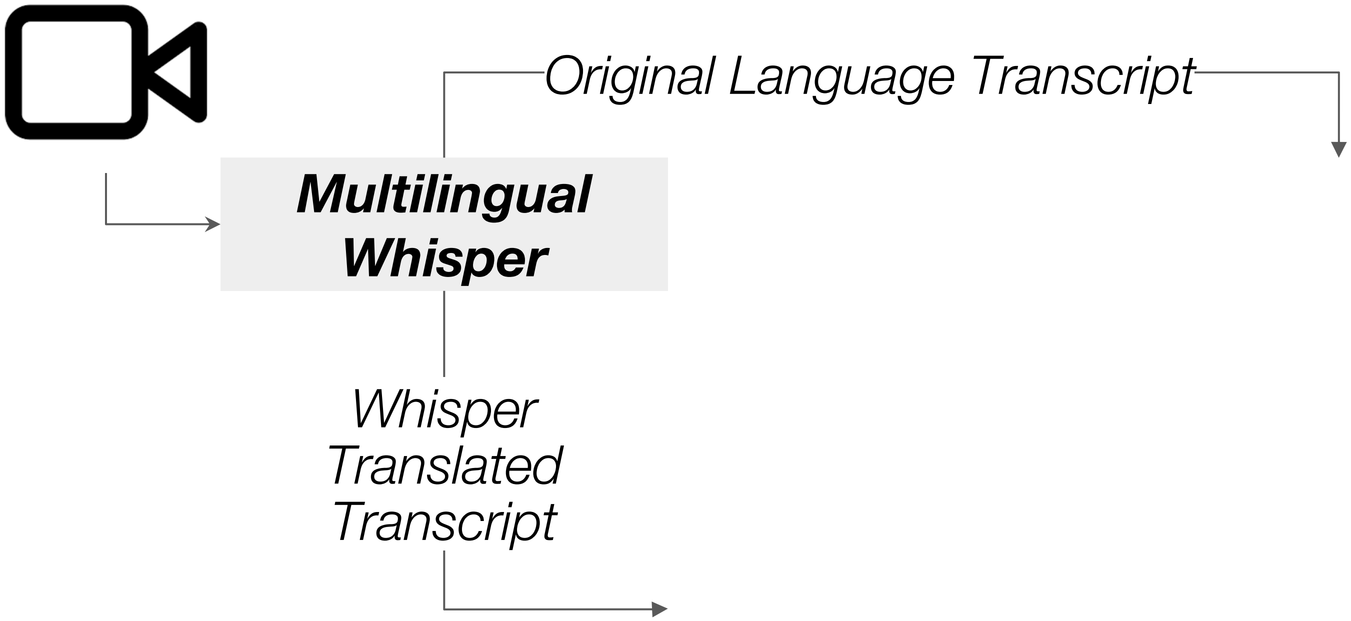
Frame Caption 16

Video Captioner

Video Caption

ASR Module

Audio Description Extraction

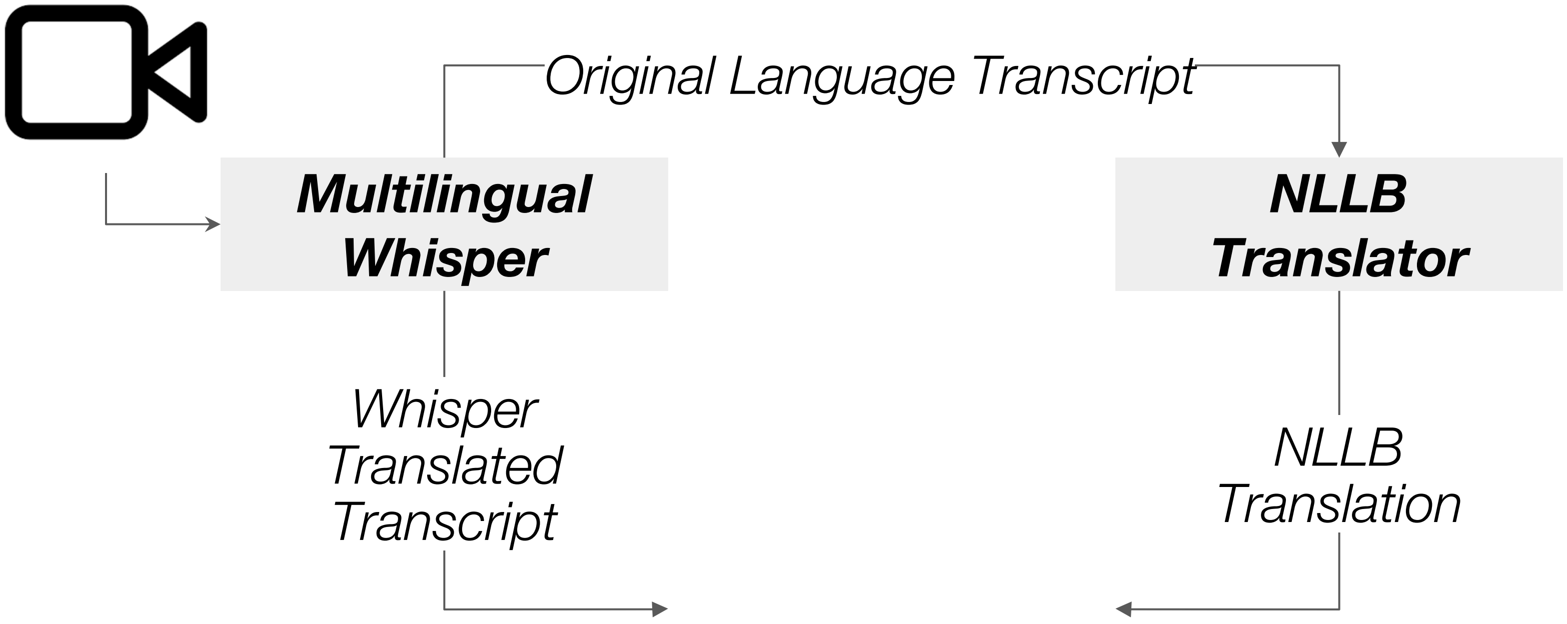


Original Language Transcript

Multilingual Whisper

Whisper Translated Transcript

Audio Description Extraction



Original Language Transcript

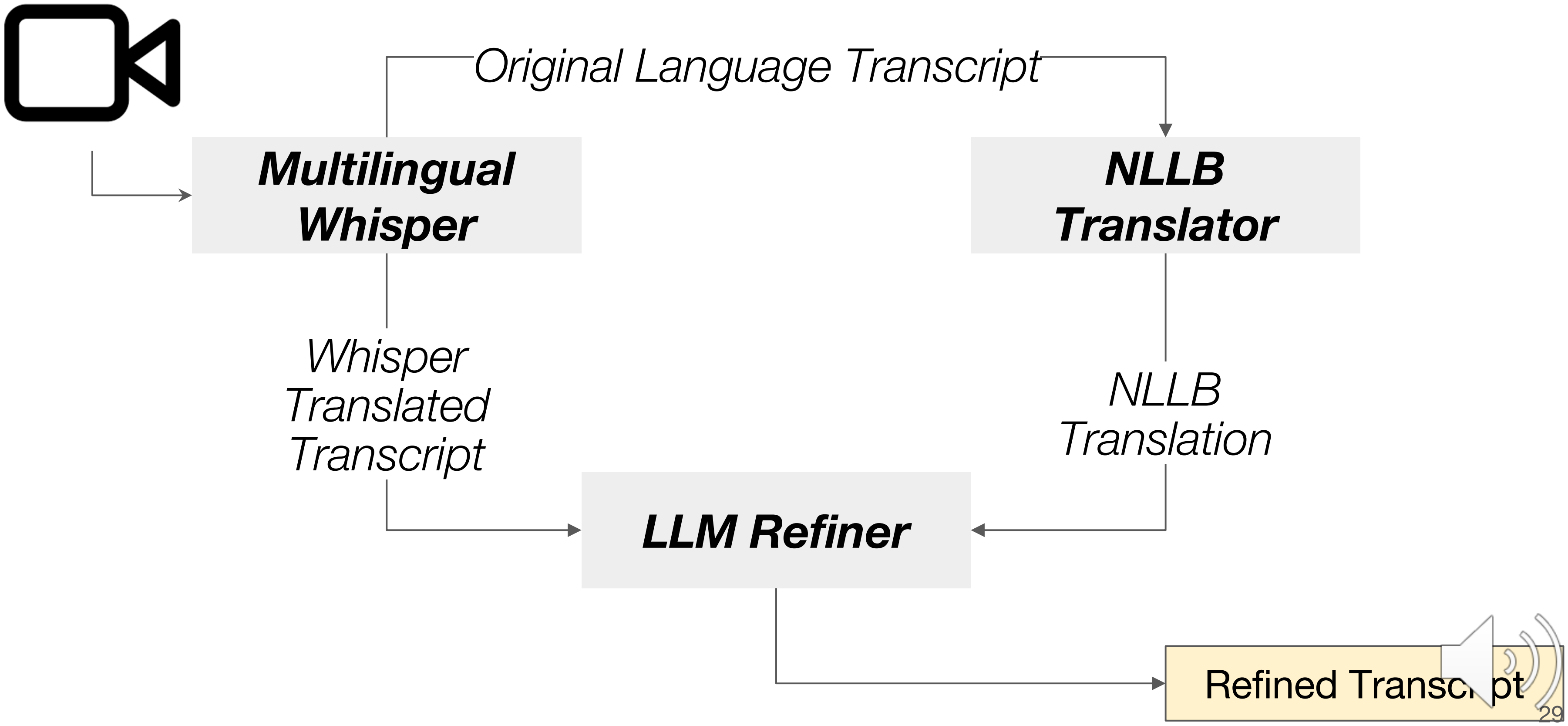
Multilingual Whisper

NLLB Translator

Whisper Translated Transcript

NLLB Translation

Audio Description Extraction



Original Language Transcript

Multilingual Whisper

NLLB Translator

Whisper Translated Transcript

NLLB Translation

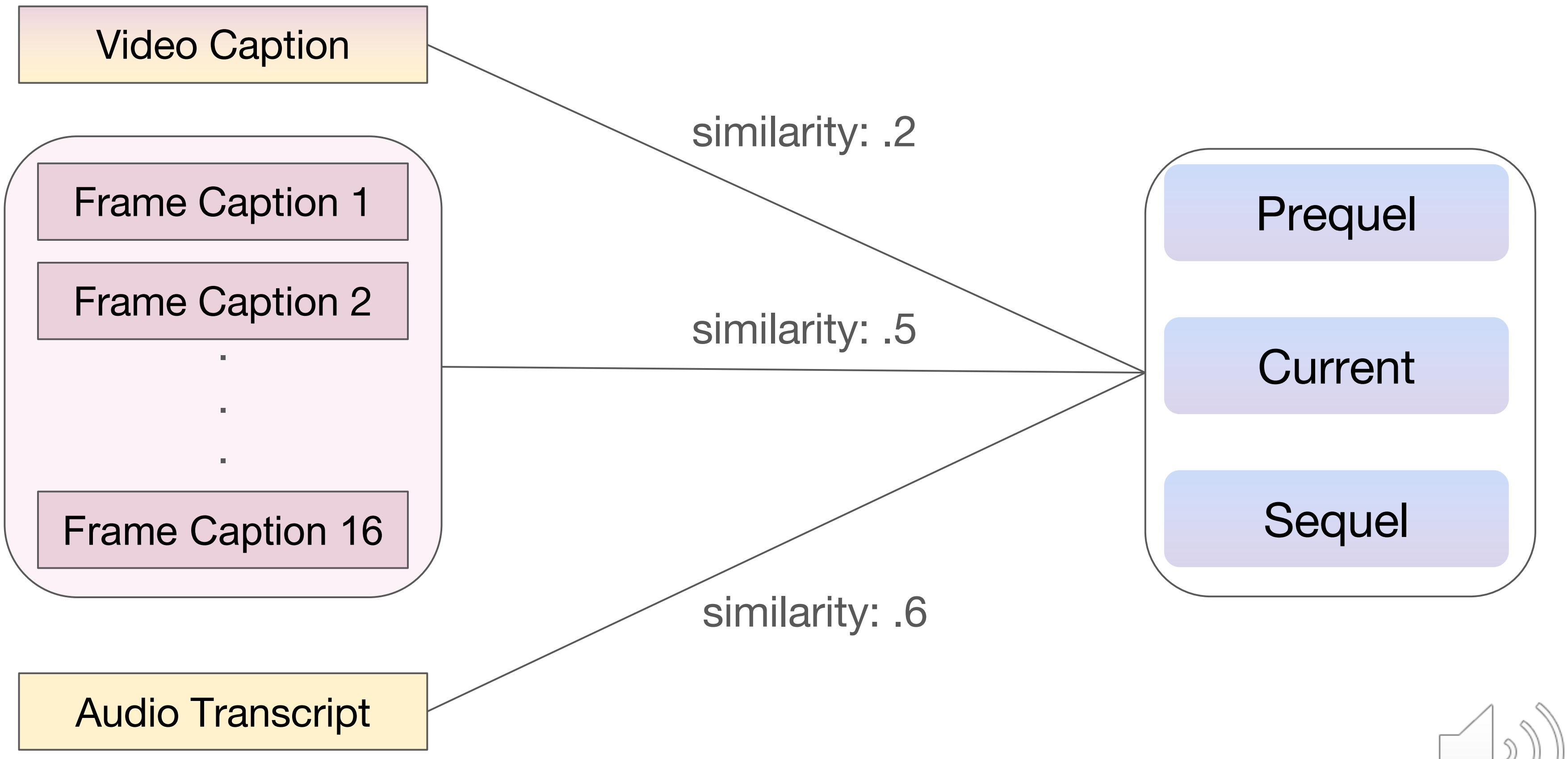
LLM Refiner

Refined Transcript

Retrieving videos is hard **without** metadata. How can we improve it…

- via enriching the query? ✅
- via information *within* the video? ✅

For a given query and potential target video…



Video Caption

Frame Caption 1

Frame Caption 2

.
.
.

Frame Caption 16

Audio Transcript

similarity: .2

similarity: .5

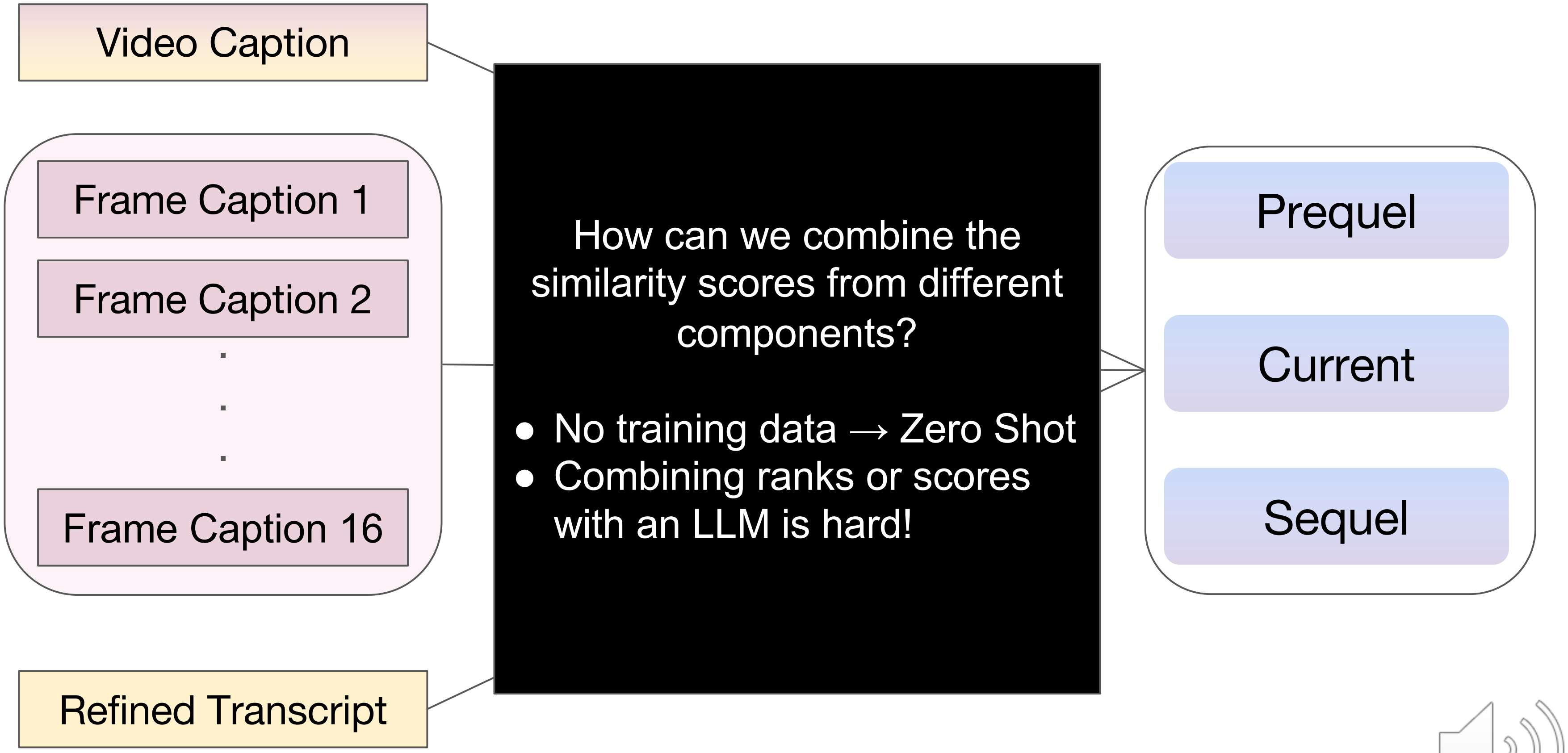
similarity: .6

Prequel

Current

Sequel

For a given query and potential target video…



Video Caption

Frame Caption 1

Frame Caption 2

.
.
.

Frame Caption 16

Refined Transcript

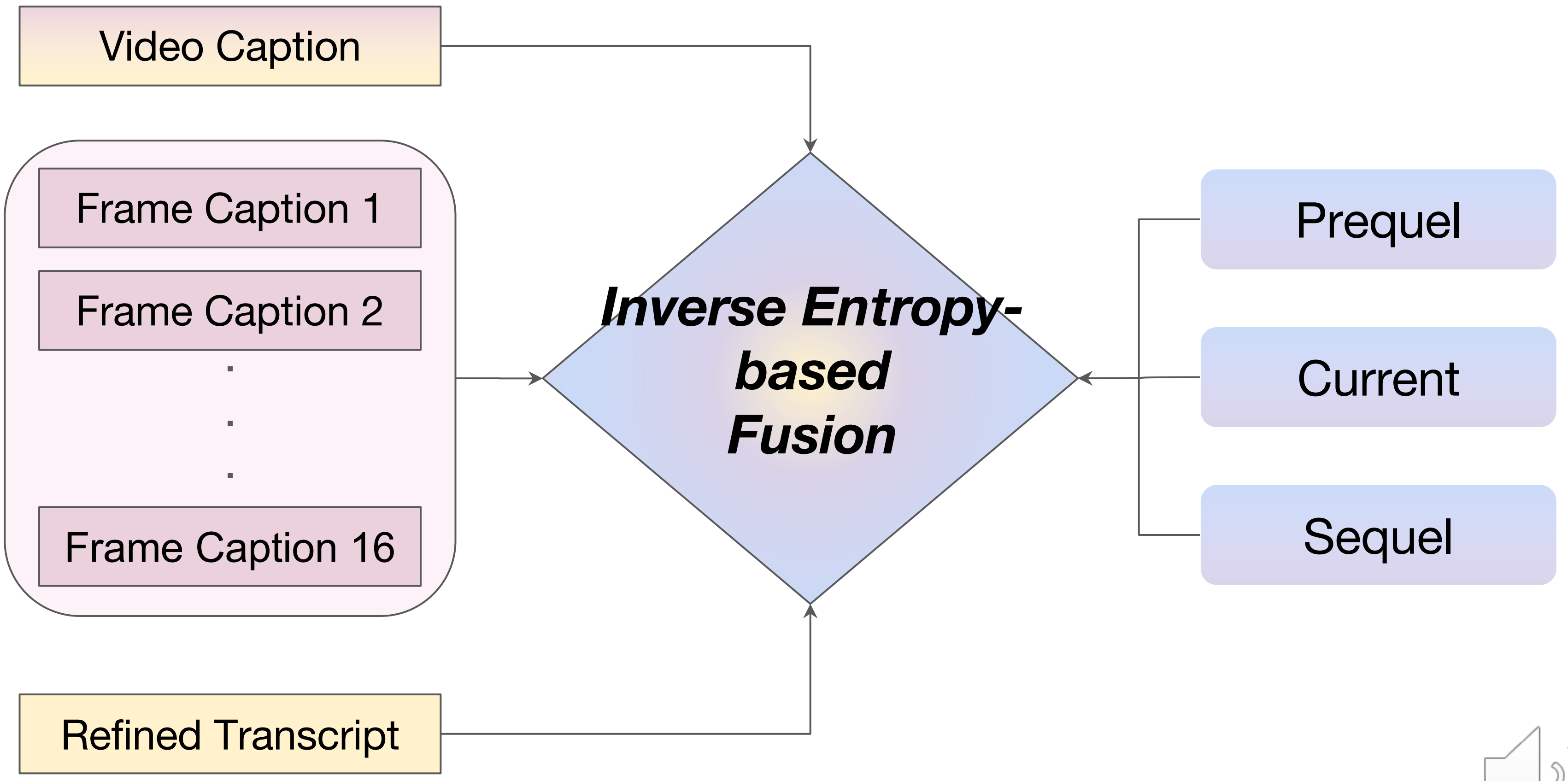
How can we combine the similarity scores from different components?

- No training data → Zero Shot
- Combining ranks or scores with an LLM is hard!

Prequel

Current

Sequel



Video Caption

Frame Caption 1

Frame Caption 2

.

.

.

Frame Caption 16

Refined Transcript

Inverse Entropy-based Fusion

Prequel

Current

Sequel

Results

MultiVENT

Number of Queries	259
Average Words in Query	27
Number of Videos	2394
Average Length of Video	83

- 5 Languages – Arabic, Chinese, English, Korean, Russian
- Event Specific dataset, i.e., “*earthquake, flood*”

MultiVENT

Model	**R@10 ↑**	**P@10 ↑**	**MRR ↑**	**NDCG ↑**
MultiCLIP	70.82	65.25	0.92	75.34
Q_2E (MultiCLIP)				
InternVideo2				
Q_2E (InternVideo2)				

MultiVENT

Model	**R@10 ↑**	**P@10 ↑**	**MRR ↑**	**NDCG ↑**
MultiCLIP	70.82	65.25	0.92	75.34
Q$_2$E (MultiCLIP)	**79.60**	**73.09**	**0.95**	**83.24**
InternVideo2				
Q$_2$E (InternVideo2)				

MultiVENT

Model	**R@10 ↑**	**P@10 ↑**	**MRR ↑**	**NDCG ↑**
MultiCLIP	70.82	65.25	0.92	75.34
Q_2E (MultiCLIP)	**79.60**	**73.09**	**0.95**	**83.24**
InternVideo2	49.12	45.44	0.68	50.45
Q_2E (InternVideo2)				

MultiVENT

Model	**R@10 ↑**	**P@10 ↑**	**MRR ↑**	**NDCG ↑**
MultiCLIP	70.82	65.25	0.92	75.34
Q_2E (MultiCLIP)	**79.60**	**73.09**	**0.95**	**83.24**
InternVideo2	49.12	45.44	0.68	50.45
Q_2E (InternVideo2)	**70.79**	**65.14**	**0.95**	**76.10**

Ablation Studies

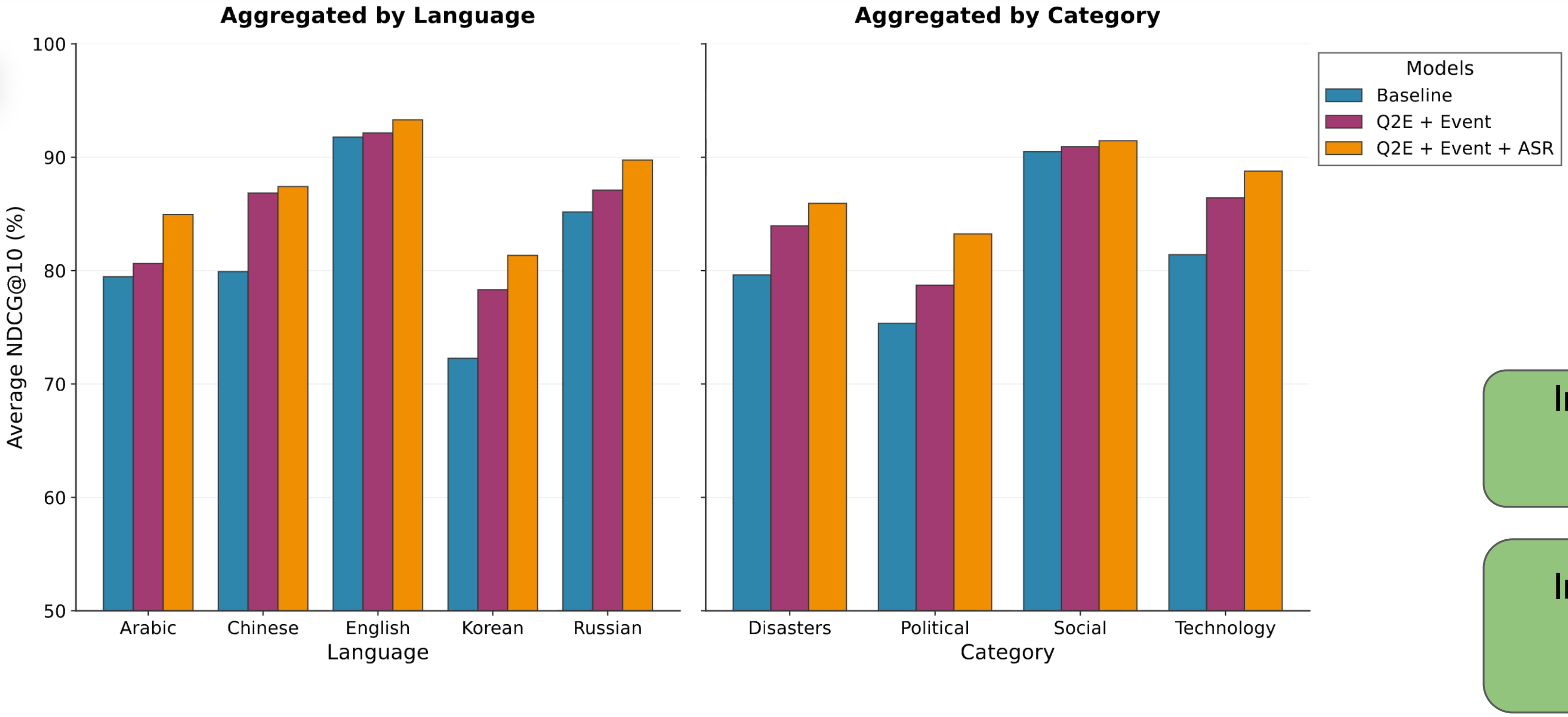
- Performance Across Different Fusion Algorithms
- Performance Across Languages
- Performance Across Categories
- Effect of Different Components
- Effect of LLM size (in paper)
- *Effect of VLM size* (in paper)
- *Effect of Key Frame Selection* (in paper)
- Effect of Rank Fusion Approaches (in paper)
- *Qualitative Examples* (in paper)

Performance Across Different **Fusion Algorithms**

Fusion Algorithm	**R@10 ↑**	**P@10 ↑**	**MRR ↑**	**NDCG ↑**
Negative Exponential Entropy	67.23	61.97	0.93	73.20
Reciprocal Rank Fusion	70.91	65.29	0.93	76.29
Maximum Aggregation	76.10	70.04	0.93	80.04
Mean Aggregation	78.64	72.47	0.95	82.44
Inverse Entropy Fusion	**79.60**	**73.09**	**0.95**	**83.24**

Inverse entropy fusion provides good reweighting

Performance Across **Language & Categories**



Improvement across all language

Improvement across all categories

Performance Of Components

Model Size	**R@10 ↑**	**P@10 ↑**	**MRR ↑**	**NDCG ↑**
Q_2E	**79.60**	**73.09**	**0.95**	**83.24**
w/o Query Sim.	78.12	71.74	0.93	81.54
w/o Video Desc.	67.74	62.43	0.93	73.96
w/o Events	77.77	71.47	0.94	81.75
Baseline	70.82	65.25	0.92	75.34

Q_2E extracts complementary information

Summary

1. We introduce **Q_2E**, a novel framework based on decomposition

2. We show that LLM's parametric knowledge can be used to enrich, otherwise, vague human queries

3. Combining both VLM and ASR gives better representation of the video

SCAN ME

Project Page

sroydip1@umbc.edu

RoyDipta.com

Auxiliary Slides

MSR-VTT-1kA

Number of Queries	995
Average Words in Query	9
Number of Videos	1000
Average Length of Video	15

- Standard 1000 test split due to High Computation
- Generic life videos, i.e., “*a man is playing with a dog*”

MSR-VTT-1kA

Model	**R@10 ↑**	**P@10 ↑**	**MRR ↑**	**NDCG ↑**
MultiCLIP	76.88	7.71	0.54	59.72
Q$_2$E (MultiCLIP)	**81.71**	**8.19**	**0.58**	**63.59**
InternVideo2	80.10	8.03	0.62	66.07
Q$_2$E (InternVideo2)	**83.72**	**8.39**	**0.65**	**69.53**

MSVD

Number of Queries	22285
Average Words in Query	8
Number of Videos	670
Average Length of Video	10

- Standard 1000 test split due to High Computation
- Generic life videos, i.e., “*A man is eating spaghetti.*”

MSVD

Model	**R@10 ↑**	**P@10 ↑**	**MRR ↑**	**NDCG ↑**
MultiCLIP	87.18	9.23	0.67	71.69
MultiCLIP + Q_2E	**89.18**	**9.45**	**0.70**	**74.10**
InternVideo2	89.47	9.49	**0.74**	77.51
InternVideo2 + Q_2E	**89.99**	**9.54**	**0.74**	**77.84**

Performance Across Different **Aggregation Methods**

Events	Captions	R@10 ↑	P@10 ↑	MRR ↑	NDCG ↑
Mean	Max	77.91	71.58	0.95	81.97
Max	Mean	0.46	0.39	0.00	0.44
Max	Mean Top 3	78.90	72.47	0.94	82.54
Max	Mean Top 5	78.91	72.47	0.94	82.50
Mean Top 3	Max	79.11	72.70	0.95	82.91
Mean Top 3	Mean Top 3	78.96	72.51	0.94	82.62
Max	Max	**79.60**	**73.09**	**0.95**	**83.24**

Q2E performs the best

Mean over captions did worst (VLM makes noisy captions)

Performance Across **LLM Size**

Model Size	**R@10 ↑**	**P@10 ↑**	**MRR ↑**	**NDCG ↑**
1B	78.71	72.28	0.95	82.50
3B	79.17	72.74	0.95	83.03
8B	79.04	72.59	0.95	82.91
70B	**79.60**	**73.09**	**0.95**	**83.24**

Q_2E provides strong multilingual improvements

Q_2E extracts complementary information

Q_2E is stable across LLM sizes

Performance Across **VLM Size**

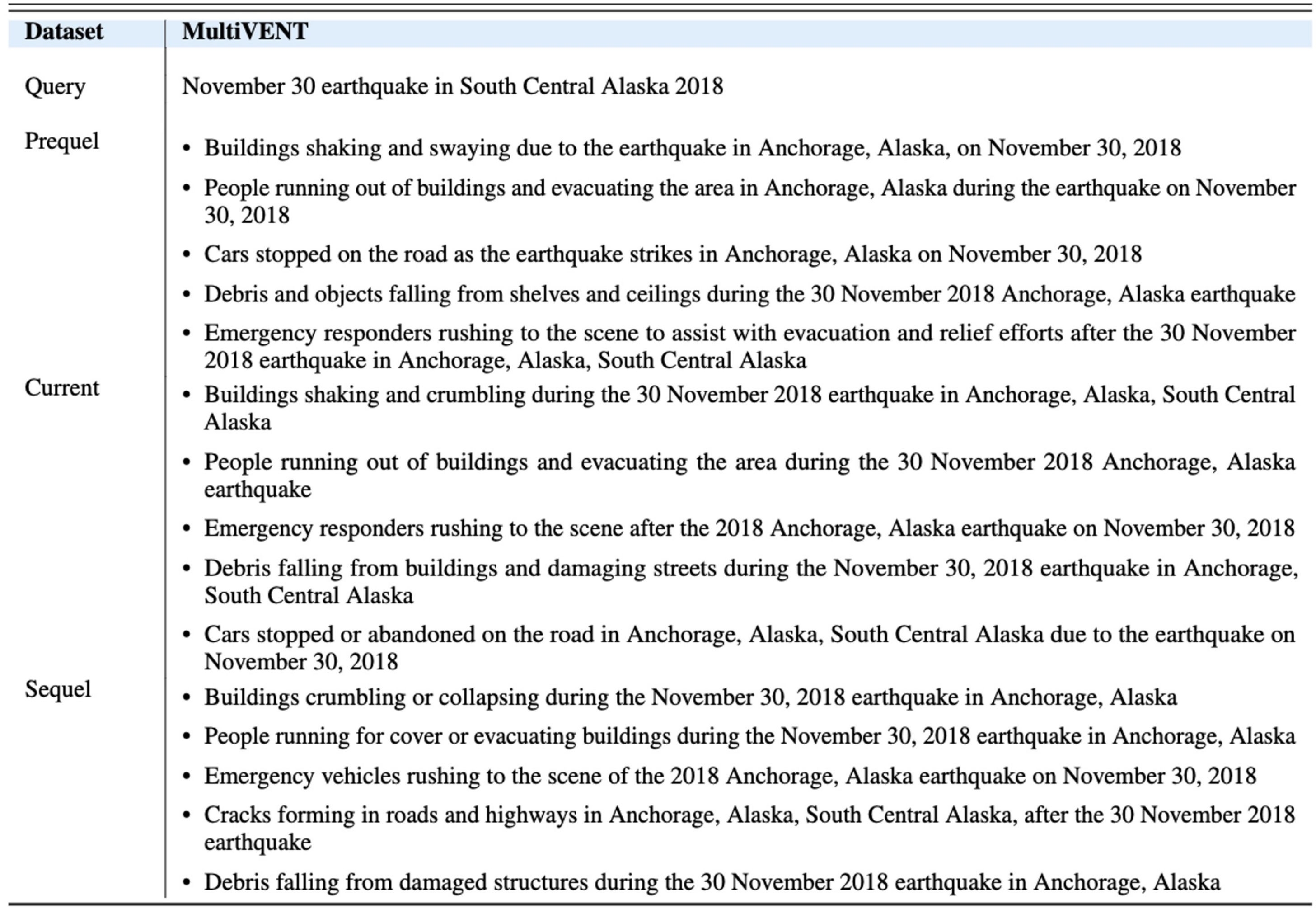
Model Size	**R@10 ↑**	**P@10 ↑**	**MRR ↑**	**NDCG ↑**
1B	71.73	66.02	0.93	75.86
2B	73.86	68.03	0.93	77.90
4B	75.85	69.92	0.93	79.81
8B	75.73	69.81	0.93	79.72
26B	**76.16**	**70.19**	**0.95**	**80.47**
38B	75.75	69.73	**0.95**	80.04

Q_2E provides strong multilingual improvements

Q_2E extracts complementary information

Q_2E is stable across LLM sizes

Q_2E is stable across big enough VLM sizes



Dataset	**MultiVENT**
Query	November 30 earthquake in South Central Alaska 2018
Prequel	• Buildings shaking and swaying due to the earthquake in Anchorage, Alaska, on November 30, 2018 • People running out of buildings and evacuating the area in Anchorage, Alaska during the earthquake on November 30, 2018 • Cars stopped on the road as the earthquake strikes in Anchorage, Alaska on November 30, 2018 • Debris and objects falling from shelves and ceilings during the 30 November 2018 Anchorage, Alaska earthquake • Emergency responders rushing to the scene to assist with evacuation and relief efforts after the 30 November 2018 earthquake in Anchorage, Alaska, South Central Alaska
Current	• Buildings shaking and crumbling during the 30 November 2018 earthquake in Anchorage, Alaska, South Central Alaska • People running out of buildings and evacuating the area during the 30 November 2018 Anchorage, Alaska earthquake • Emergency responders rushing to the scene after the 2018 Anchorage, Alaska earthquake on November 30, 2018 • Debris falling from buildings and damaging streets during the November 30, 2018 earthquake in Anchorage, South Central Alaska • Cars stopped or abandoned on the road in Anchorage, Alaska, South Central Alaska due to the earthquake on November 30, 2018
Sequel	• Buildings crumbling or collapsing during the November 30, 2018 earthquake in Anchorage, Alaska • People running for cover or evacuating buildings during the November 30, 2018 earthquake in Anchorage, Alaska • Emergency vehicles rushing to the scene of the 2018 Anchorage, Alaska earthquake on November 30, 2018 • Cracks forming in roads and highways in Anchorage, Alaska, South Central Alaska, after the 30 November 2018 earthquake • Debris falling from damaged structures during the 30 November 2018 earthquake in Anchorage, Alaska