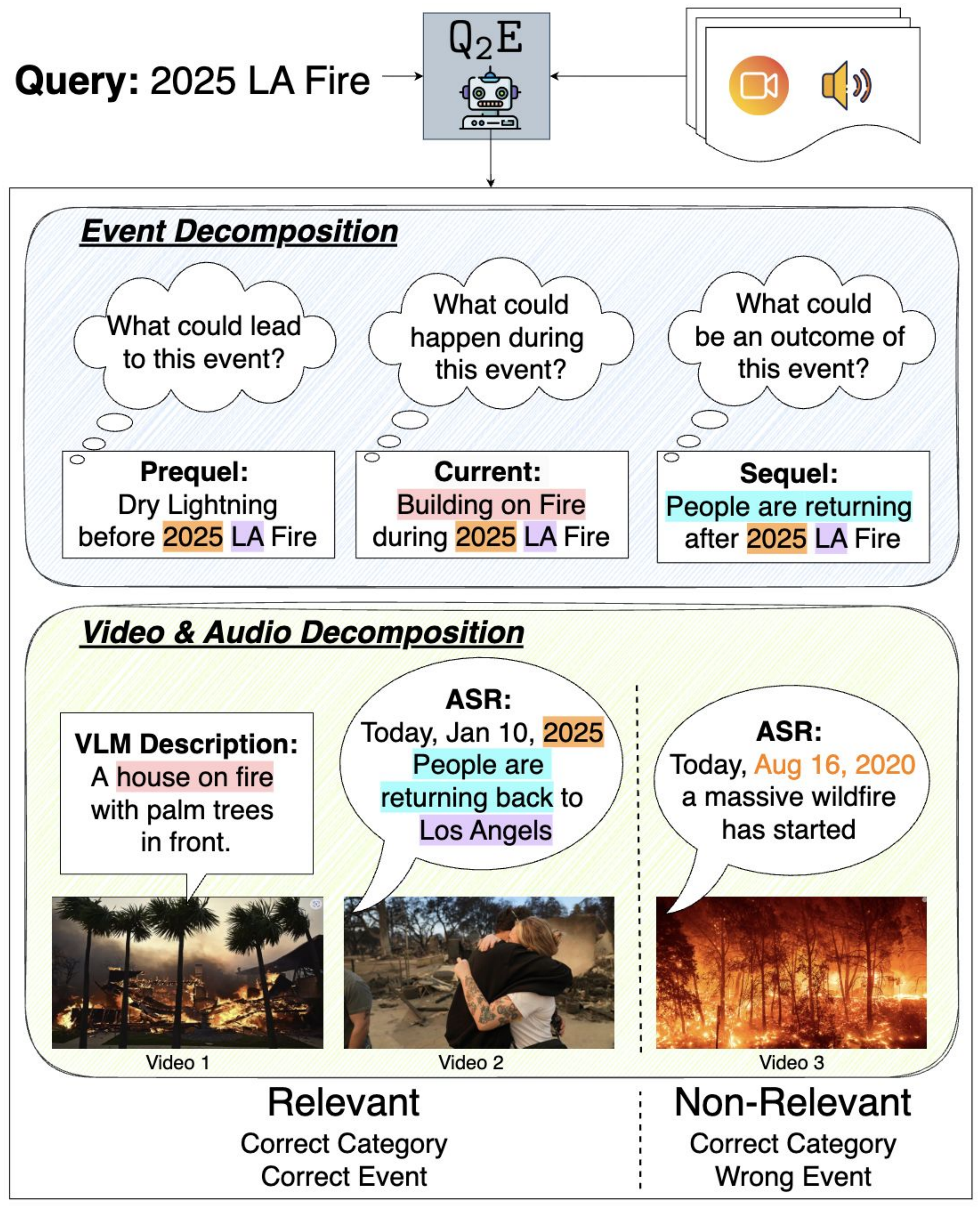
Q_2E: Query-to-Event Decomposition for Zero-Shot Multilingual Text-to-Video Retrieval

Shubhashis Roy Dipta and Francis Ferraro, University of Maryland, Baltimore County.

🤔🤔 Why does Retrieval Fail on Complex Events **WITHOUT** Metadata? 🤔🤔


Query: 2025 LA Fire
Q_2E
Event Decomposition
What could lead to this event?
What could happen during this event?
What could be an outcome of this event?
Prequel:
Dry Lightning before 2025 LA Fire
Current:
Building on Fire during 2025 LA Fire
Sequel:
People are returning after 2025 LA Fire
Video & Audio Decomposition
VLM Description:
A house on fire with palm trees in front.
ASR:
Today, Jan 10, 2025 People are returning back to Los Angels
ASR:
Today, Aug 16, 2020 a massive wildfire has started
Video 1
Video 2
Video 3
Relevant
Correct Category
Correct Event
Non-Relevant
Correct Category
Wrong Event

Problem:

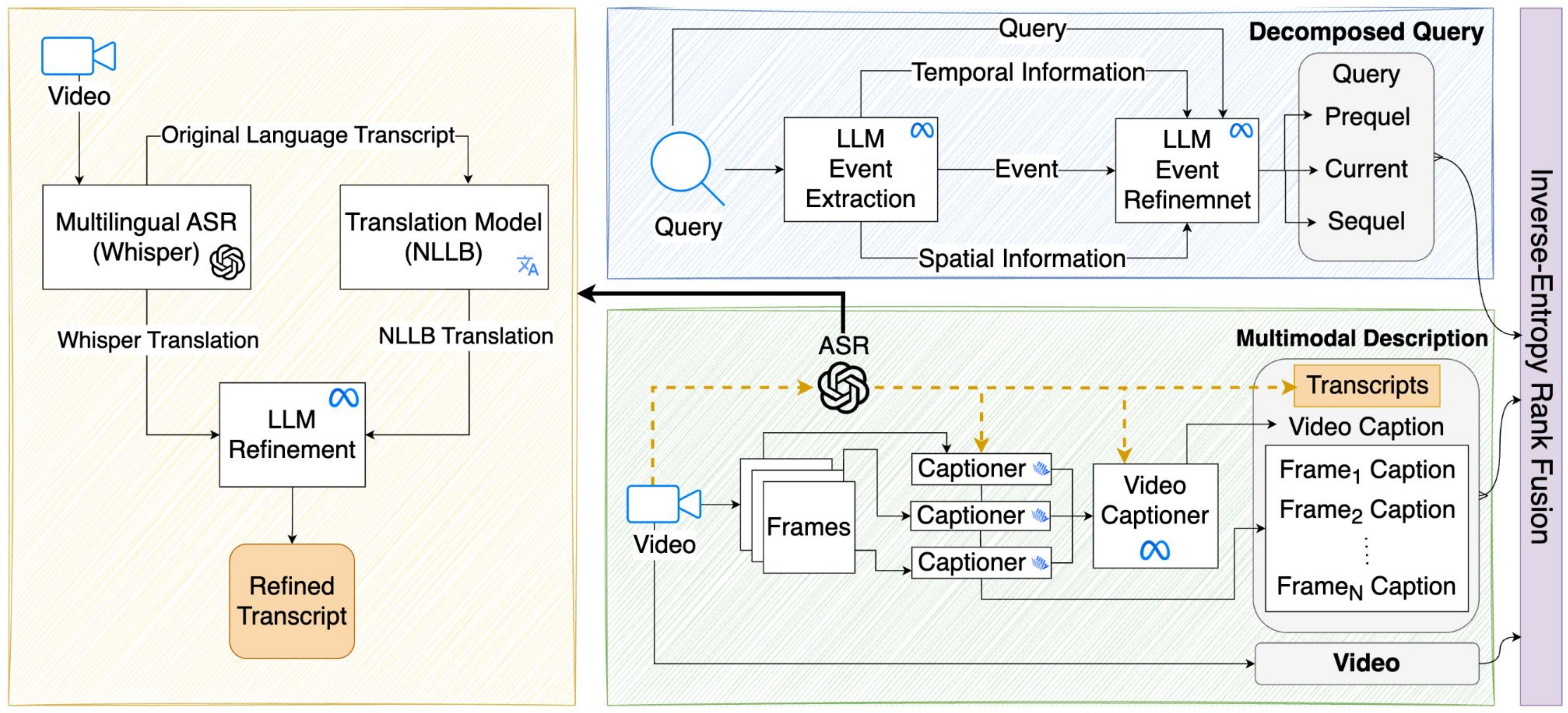
- ➢ Metadata is not always present
- ➢ Users query complex events (e.g., "2025 LA Fire"), but standard embedding models **only look for generic semantic matches**

Gap:

- ➢ A single embedding cannot capture the nuances of a real-world event

Contribution:

- ➢ A novel framework to enrich query and videos using prior knowledge from LLM
- ➢ LLM's parametric knowledge can be used to enrich, otherwise, vague human queries
- ➢ Combining both VLM and ASR gives better representation of the video



Query Decomposition

1. Temporal Decomposition
2. Entity Extractions
3. Refine to Natural Query

Video Decomposition

1. Frame Captioning
2. ASR Transcript
3. Holistic Video Summaries

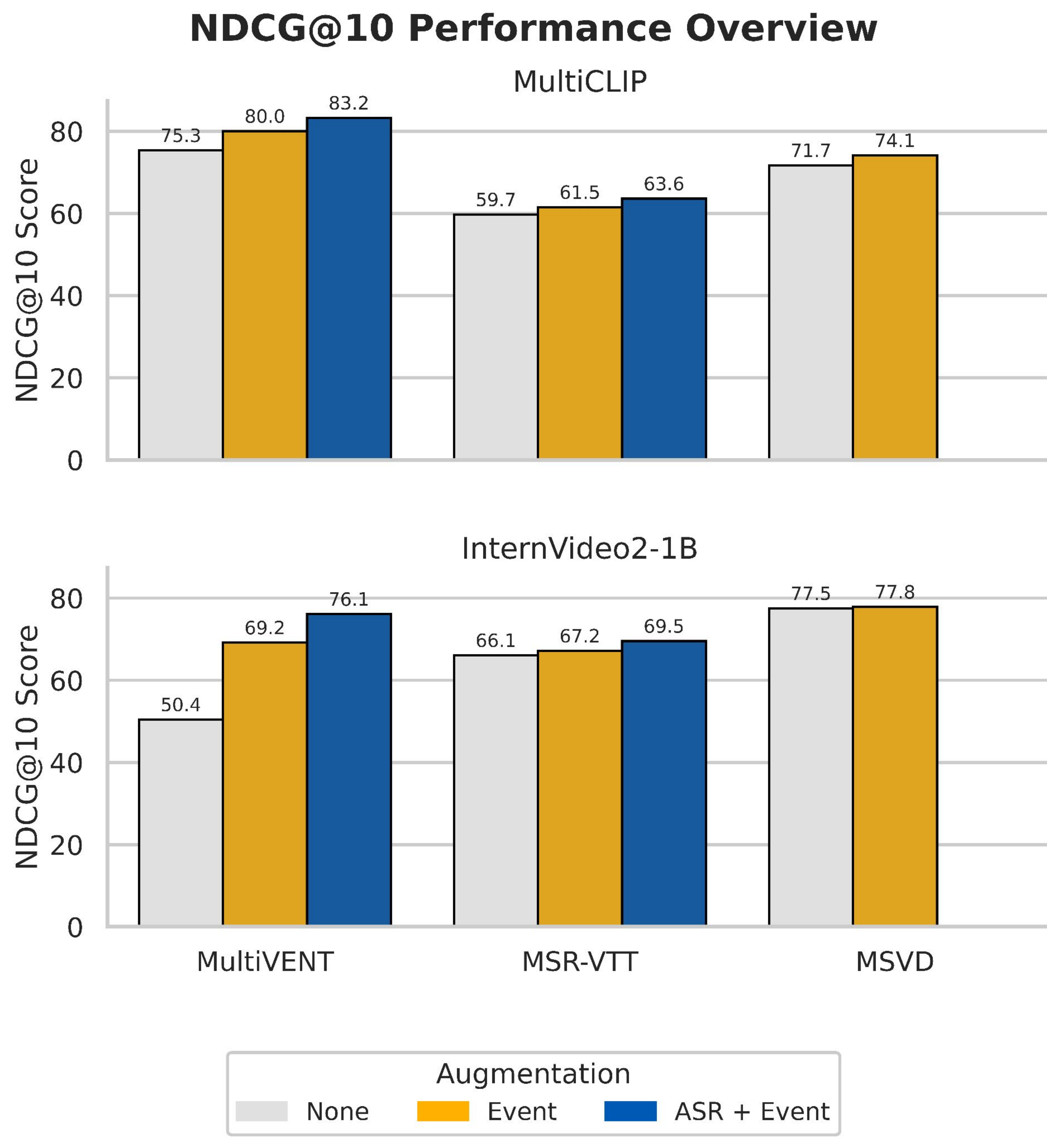
Audio Decomposition

1. Whisper Translate & Transcribe
2. NLLB Translate
3. LLM Refinement

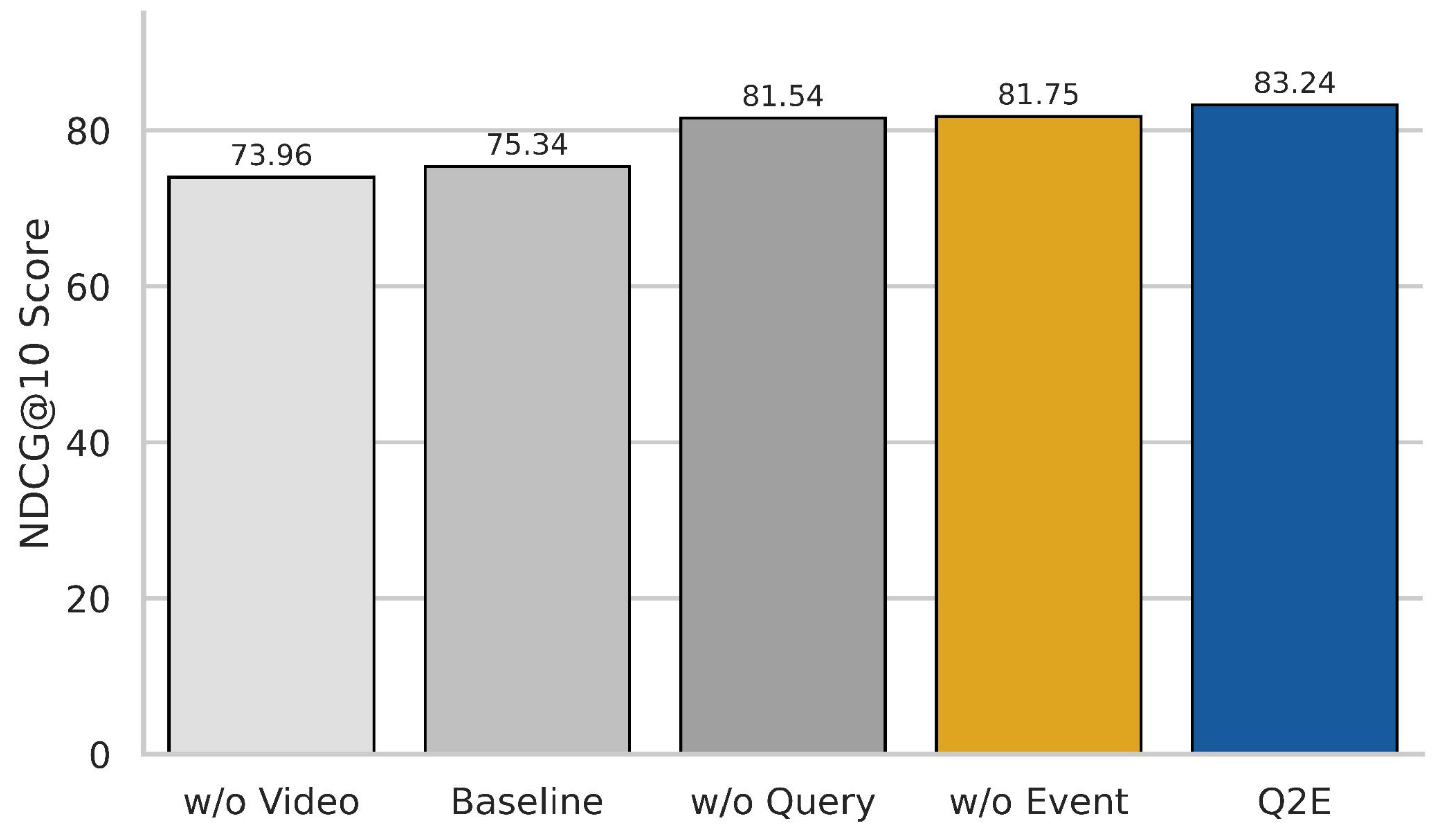
Scoring & Score Fusion

1. Video & Text Alignment
2. Global Maximum
3. Entropy-Based Fusion

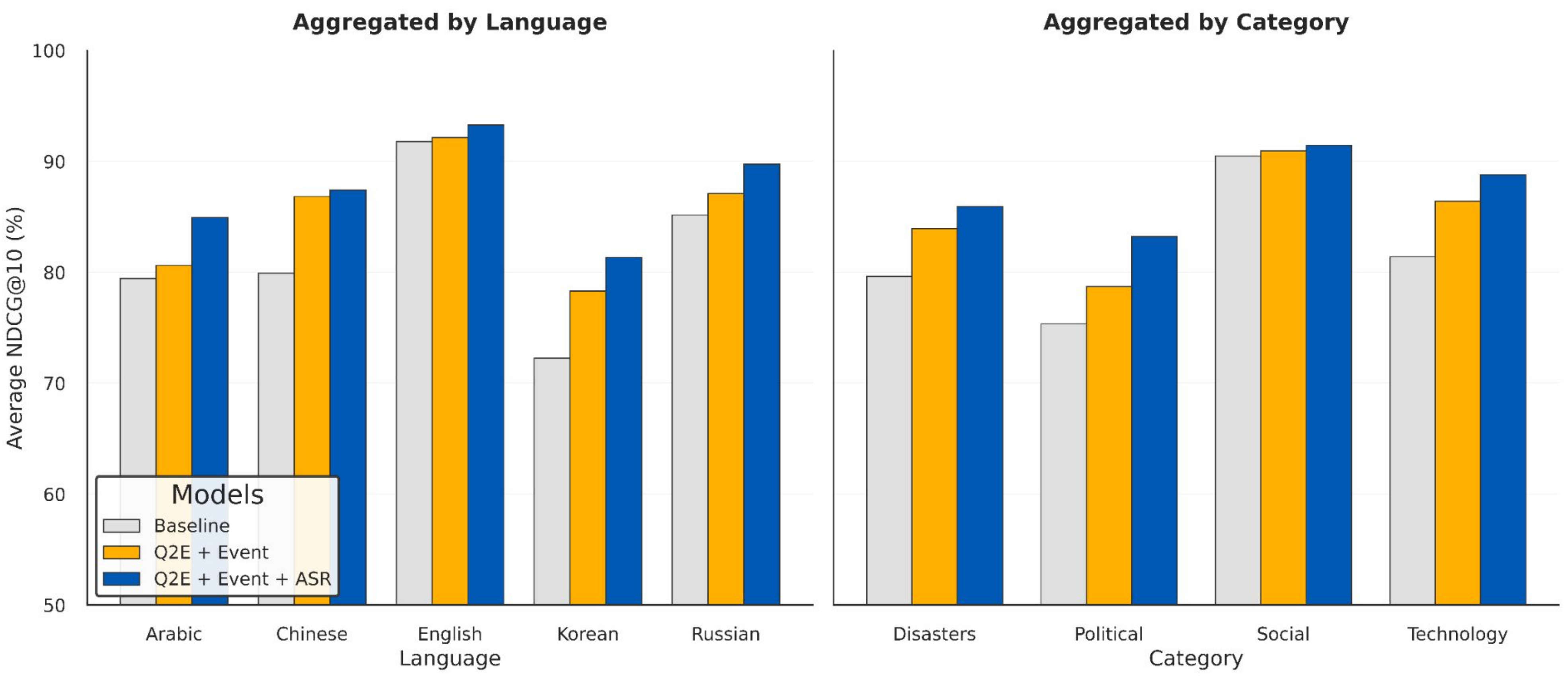
Main Results



Takeaway 2: Q_2E Extracts Complementary Information



Takeaway 3: Consistent Improvement Across Language & Categories


Aggregated by Language
Average NDCG@10 (%)
Models
Baseline
Q2E + Event
Q2E + Event + ASR
Arabic
Chinese
English
Korean
Russian
Language
Aggregated by Category
Disasters
Political
Social
Technology
Category