NEURAL INFORMATION PROCESSING SYSTEMS 2026

Driving Video Retrieval for Complex Queries with Structured Grounding


Manyi Yao†


Sparsh Garg‡


Christian Shelton†

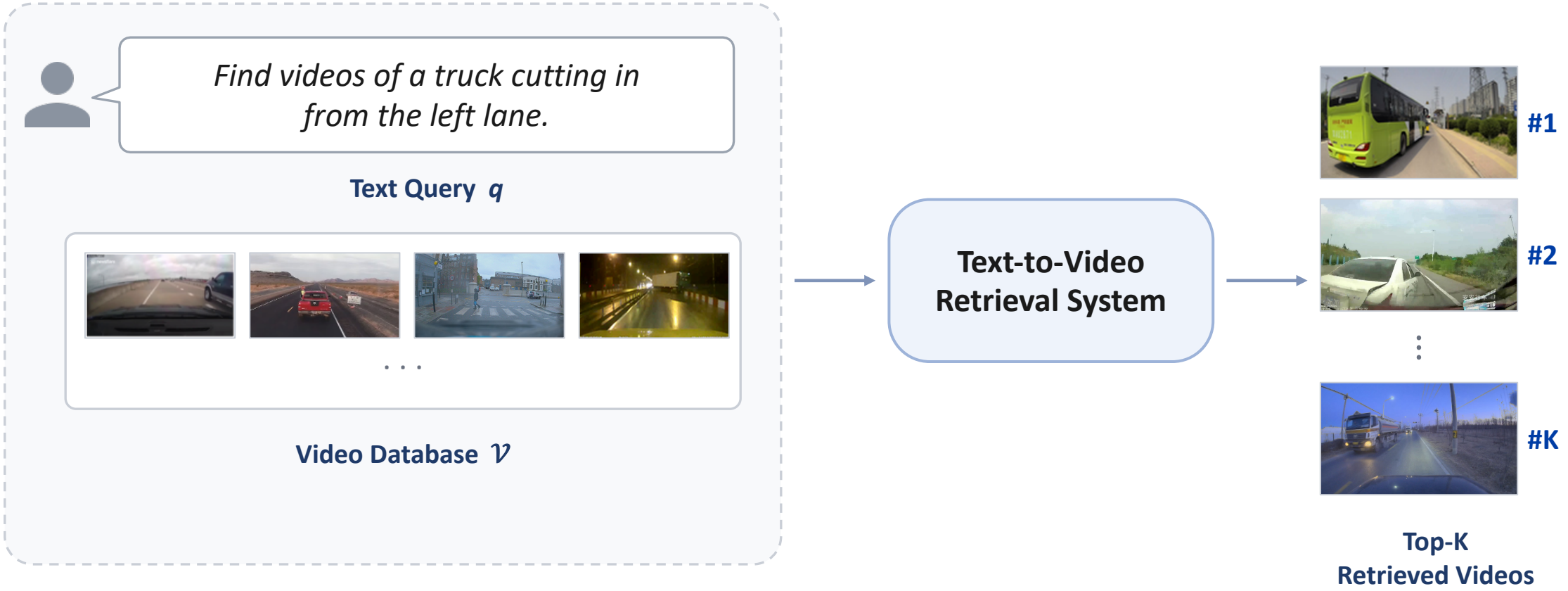

Amit Roy-Chowdhury†


Abhishek Aich‡

† UC RIVERSIDE

‡ NEC LABORATORIES AMERICA

Driving Video Retrieval



We want to find **rare, safety-critical events**, like cut-ins and braking sharply, in large video collections.

Classic Retrieval Approaches

- **Embedding-based (dense retrieval)**: encode the query and each video as vectors; rank by cosine similarity.

  e.g., CLIP[1], SigLIP2[2], Qwen3-VL-Embedding[3]

- **Lexical (sparse retrieval)**: match query keywords against text descriptions of the videos.

  e.g., TF-IDF[4], BM25[5]

- **Neuro-symbolic (structured grounding)**: turn the query into a program or logic formula and check it on perception outputs.

  e.g., ViperGPT[6], NSVS-TL[7]

[1] Radford et al., Learning Transferable Visual Models From Natural Language Supervision, ICML 2021.
[2] Tschannen et al., SigLIP 2: Multilingual Vision-Language Encoders, arXiv 2025.
[3] Li et al., Qwen3-VL-Embedding and Qwen3-VL-Reranker, arXiv 2026.
[4] Sparck Jones, A Statistical Interpretation of Term Specificity, Journal of Documentation 1972.
[5] Robertson et al., The Probabilistic Relevance Framework: BM25 and Beyond, FnTIR 2009.
[6] Surís et al., ViperGPT: Visual Inference via Python Execution for Reasoning, ICCV 2023.
[7] Choi et al., Towards Neuro-Symbolic Video Understanding, ECCV 2024.

The Gap of Embedding-Based Retrieval

Query: *Find videos where a vehicle **changes lane** in front of the ego vehicle.*

Result 1



Car stays in its lane.

Result 2



Cars cross an intersection.

Result 3



Vehicle in the opposite lane.

✓ Captures **what's in the scene**.

✗ Fails on **fine-grained motion**. **Failure mode 1: Dilution**

Symbolic Rules: Events as Executable Filters



Per-frame structured records[1] j_v

```
{
  frame_index: 30,
  objects: [
    { id: 3, class: car,
      distance: 7, loc_x: -1.6,
      lane_rel: 0 },
    ...
  ]
}
```

illustrative values

```
{
  entity:
    class_filter: [car, truck]
    agg_mode: any

  hard_gates:
    distance < 40.0

  soft_scores:
    loc_x  max_change
    window_seconds: 4.0
    magnitude: 14.0
}
```

Entity selector
who

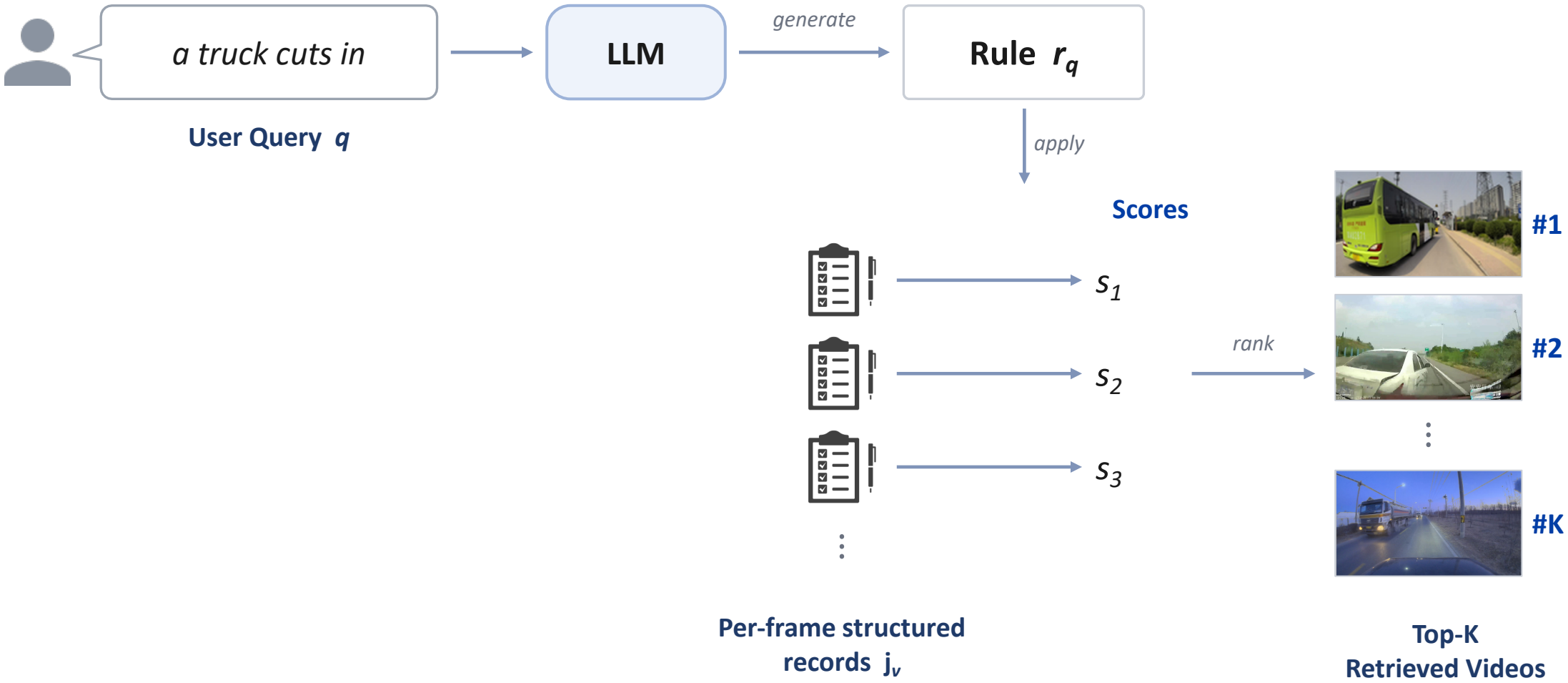
Hard gate
what must hold

Soft score
how strongly

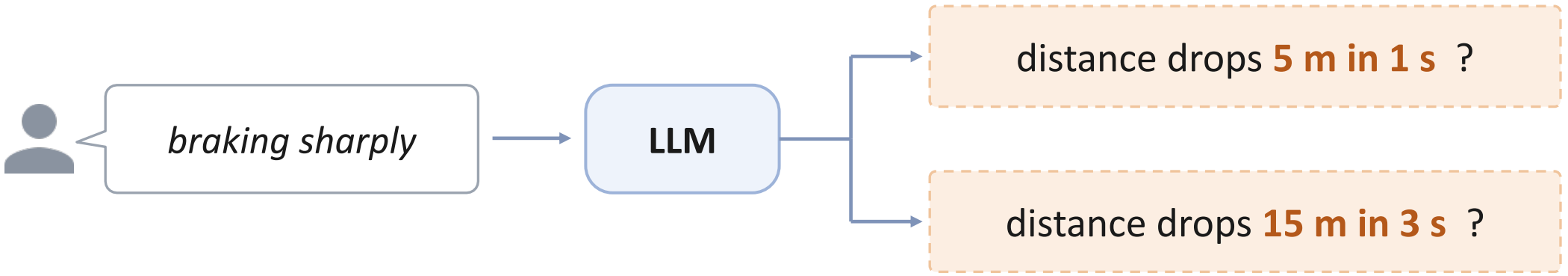
Symbolic rule r for "cut-in"

[1] Yao et al., iFinder: Structured Zero-Shot LLM Grounding for Dash-Cam Video Reasoning, NeurIPS 2025.

Retrieval with Symbolic Rules



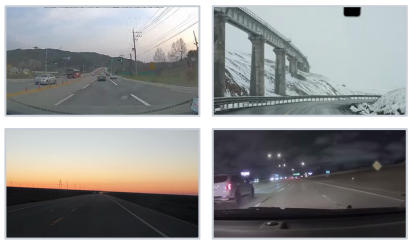
Why Not Just Use LLM-Generated Rules?



LLM-generated rules are structurally plausible but **miscalibrated in thresholds**.

Failure mode 2: Miscalibration

Event Mining: From Captions to Weak Labels



Auxiliary set $\mathcal{V}'$

unlabeled, disjoint from all benchmarks

Video captioner

"A white car moves from the left lane into the ego lane ahead …"

Captions

LLM

extract

Event mentions

- (cut-in, car)
- (cutting in, truck)
- (cut-out, car)

merge

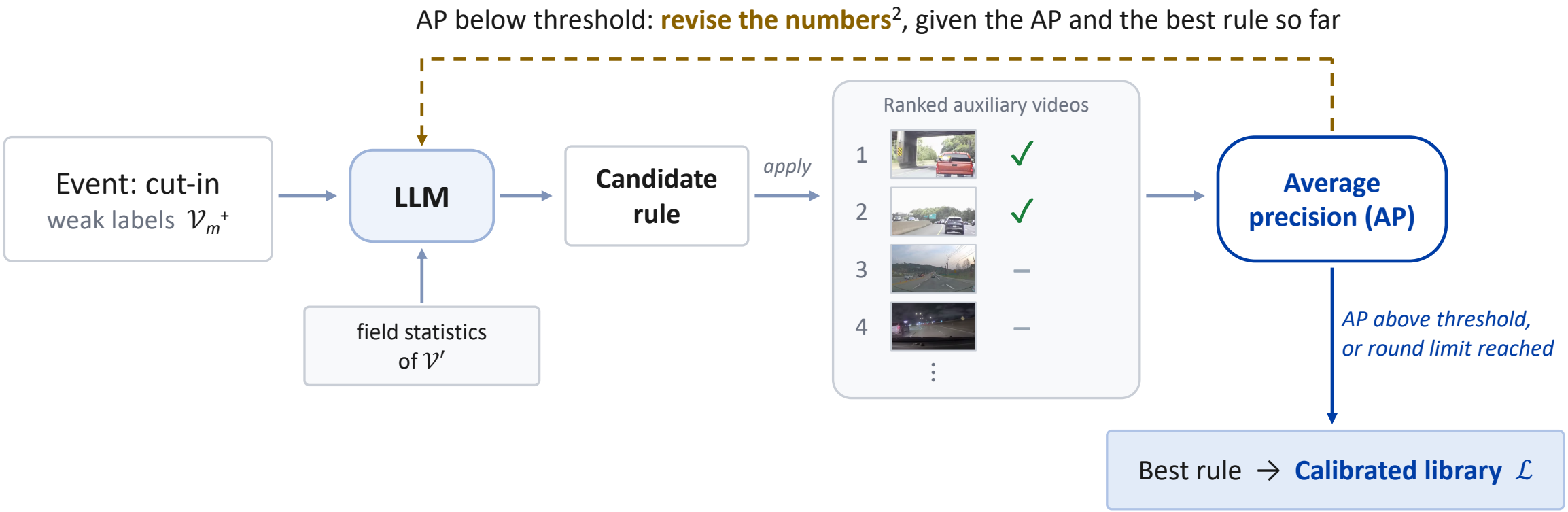
cut-in	$\mathcal{V}_m^+$:	  …
cut-out	$\mathcal{V}_m^+$:	  …

Events $\mathcal{E}$ with weak labels

caption mentions e_m → $\mathcal{V}_m^+$, otherwise $\mathcal{V}_m^-$

Captions are used only to label videos. The rules are scored on the **structured records**.

Calibration via Learning-to-Rank[1]



AP is high when the weak positives (✓) are ranked near the top.

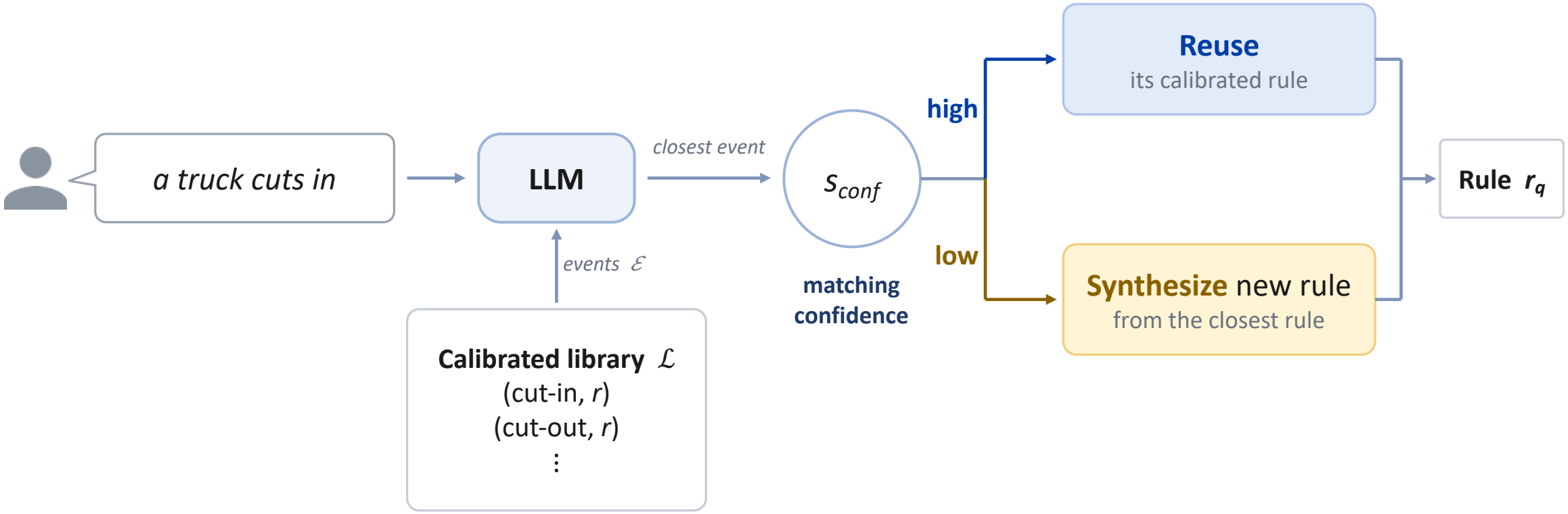
Each round keeps the rule **structure** and changes only its **numbers**.

[1] Liu, Learning to Rank for Information Retrieval, FnTIR 2009.

[2] Yang et al., Large Language Models as Optimizers, ICLR 2024.

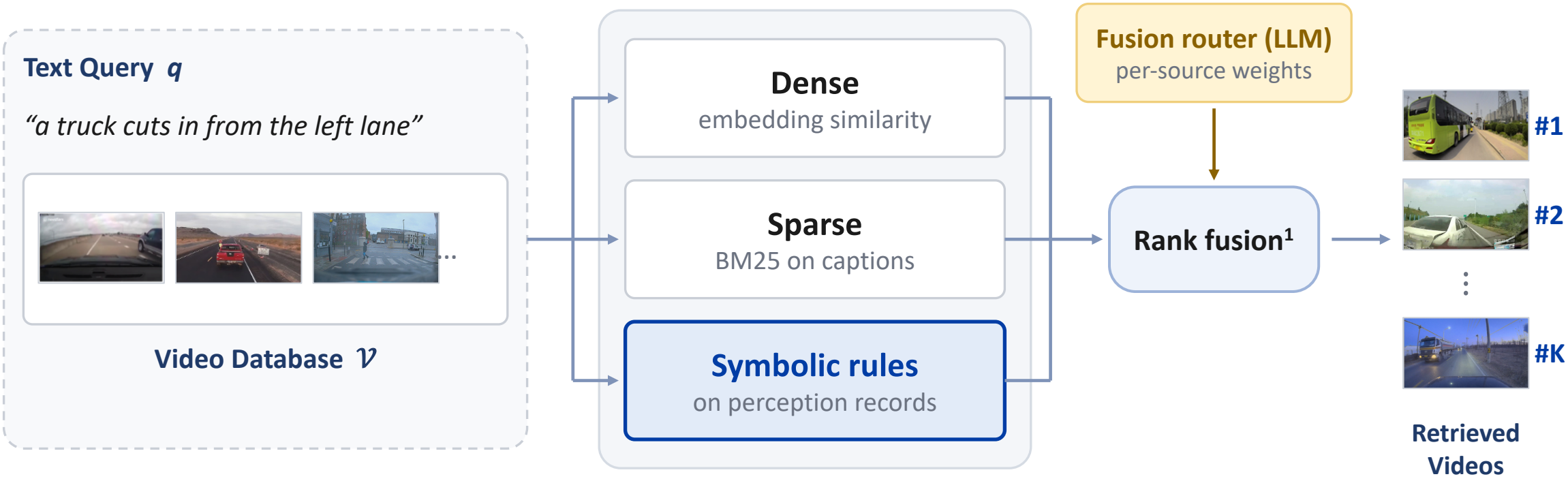
Adaptation at Retrieval Time

A library covers a fixed set of events, but a user can ask for **any** event. **Failure mode 3: Coverage**



New rules inherit the **calibrated thresholds** of the closest event.

Multi-Source Fusion



The rules **complement** dense and sparse retrieval, which cover descriptive content and rare objects.

[1] Cormack et al., Reciprocal Rank Fusion, SIGIR 2009.

Datasets

DrivingDojo[1]

583 videos, 41 motion events

motion-event annotations by us, released

Example: *a bus cuts in*

MM-AU[2]

1,953 test videos, 49 accident types

Example: *ego car skids on a snowy road*

CarCrashDataset[3]

110 videos, 45 accident causes

Example: *a motorcycle crash in city traffic*



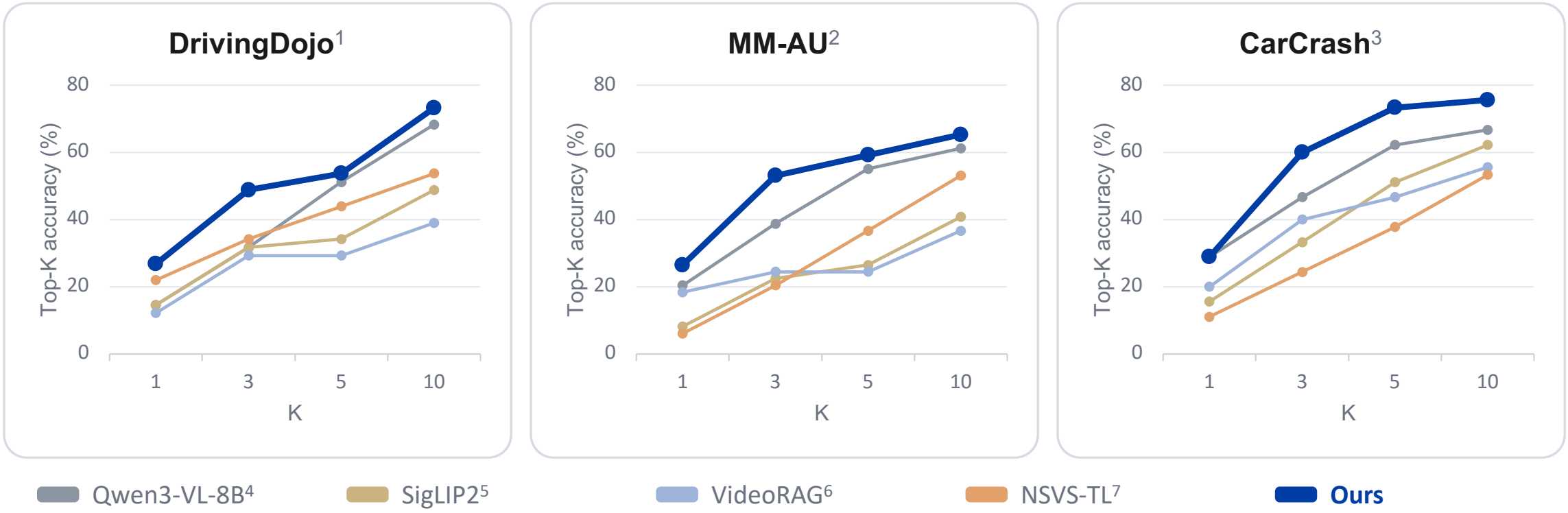
Three frames from one video of each dataset

[1] Wang et al., DrivingDojo Dataset, NeurIPS D&B 2024.

[2] Fang et al., Abductive Ego-View Accident Video Understanding, CVPR 2024.

[3] Bao et al., Uncertainty-based Traffic Accident Anticipation, ACM MM 2020.

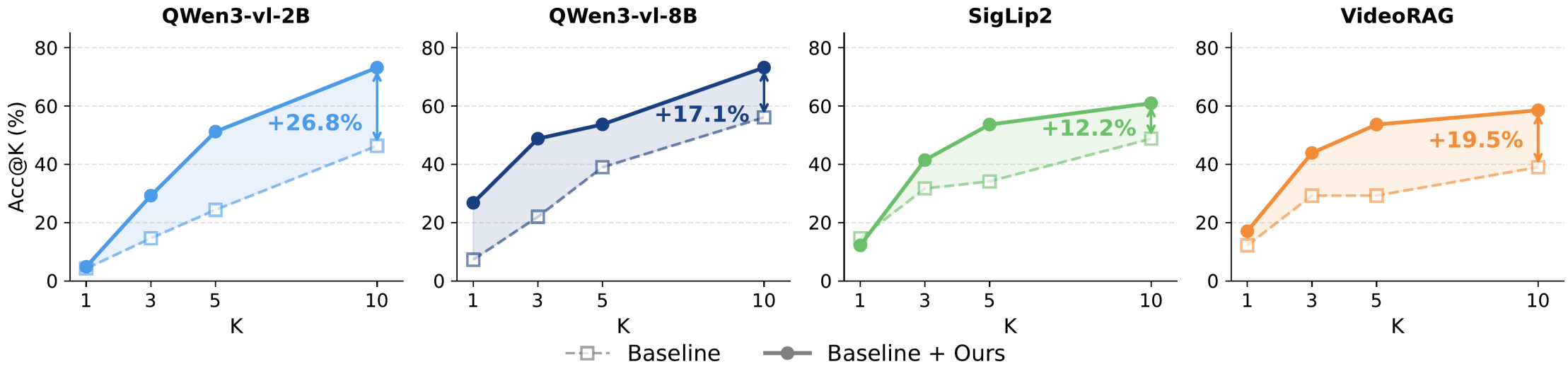
Main Results



Ours outperforms all baselines at every K in top-K accuracy.

[1] Wang et al., DrivingDojo Dataset, NeurIPS D&B 2024.
[2] Fang et al., Abductive Ego-View Accident Video Understanding, CVPR 2024.
[3] Bao et al., Uncertainty-based Traffic Accident Anticipation, ACM MM 2020.
[4] Li et al., Qwen3-VL-Embedding and Qwen3-VL-Reranker, arXiv 2026.
[5] Tschannen et al., SigLIP 2: Multilingual Vision-Language Encoders, arXiv 2025.
[6] Jeong et al., VideoRAG: Retrieval-Augmented Generation over Video Corpus, Findings of ACL 2025.
[7] Choi et al., Towards Neuro-Symbolic Video Understanding, ECCV 2024.

Dilution: Plug-in Generalizability

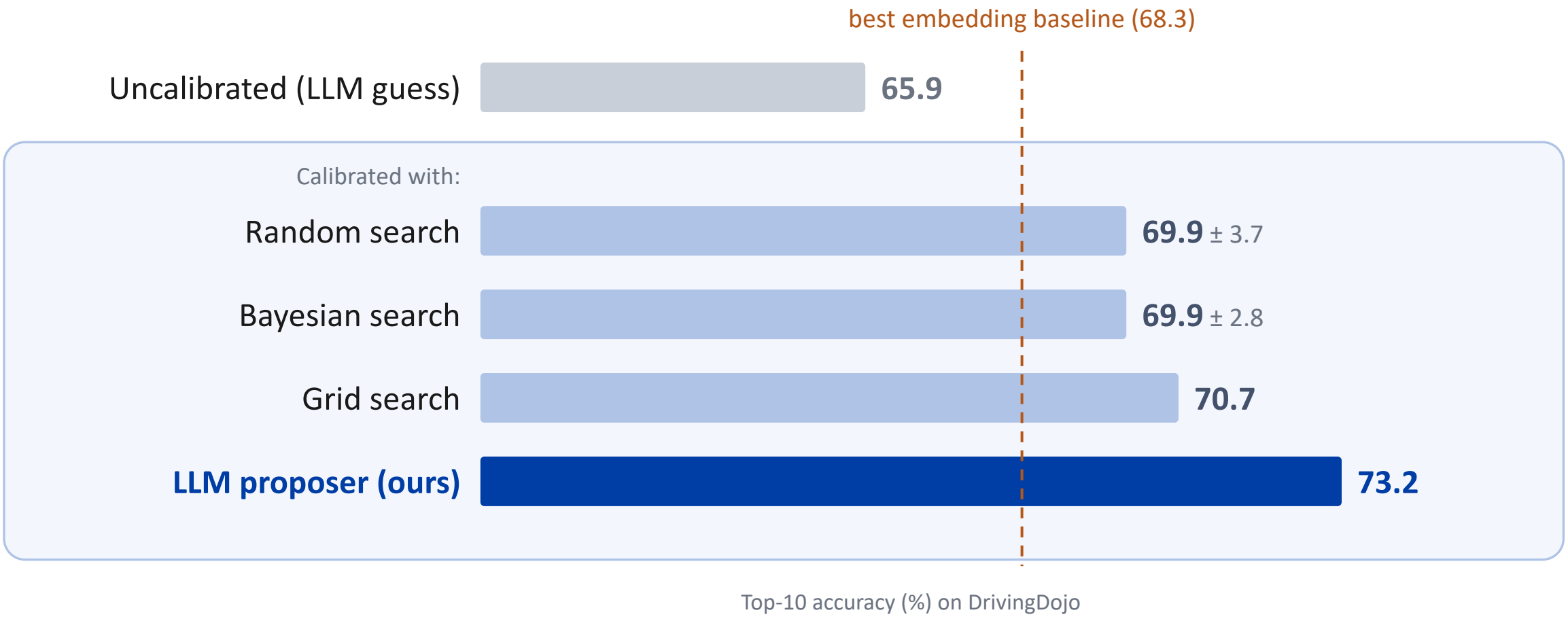


Top-10 accuracy on DrivingDojo[1]

Our method seamlessly plugs into any dense retriever and consistently boosts performance.

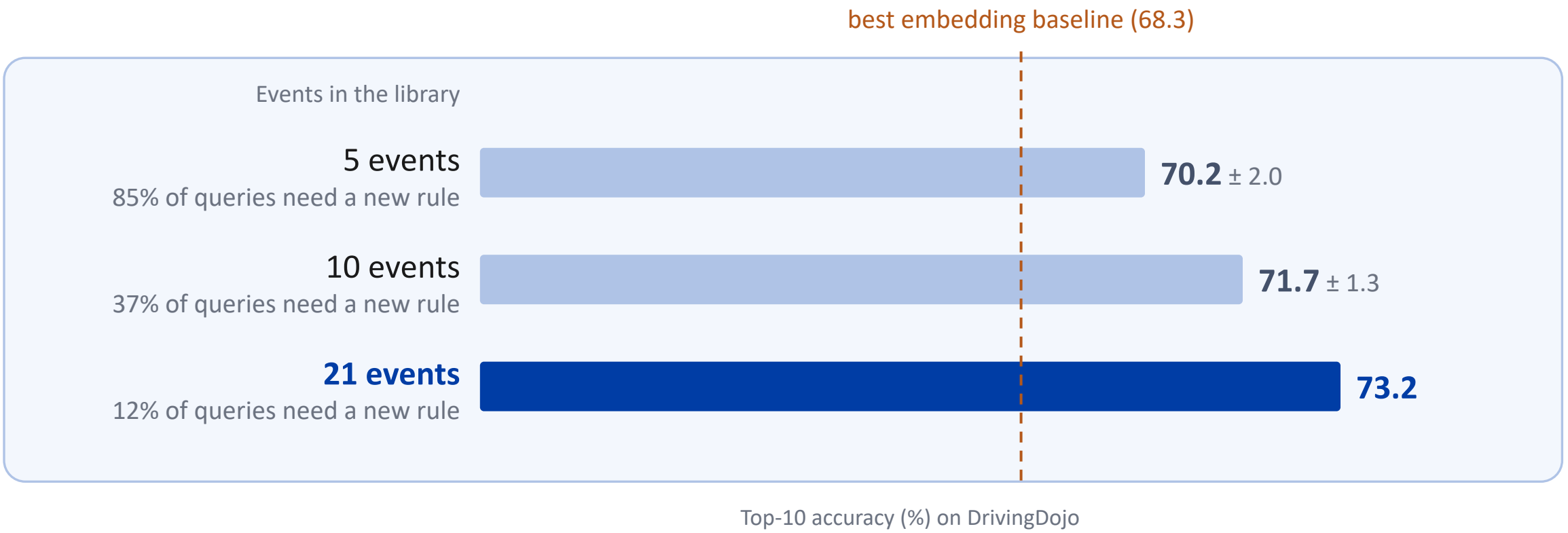
[1] Wang et al., DrivingDojo Dataset, NeurIPS D&B 2024.

Miscalibration: Calibration Proposer Comparison



Our LLM proposer outperforms traditional search methods in calibrating symbolic rules.

Coverage: Library Size and Query-Time Adaptation



Performance scales with library size, while query-time adaptation keeps small libraries strong.

Summary

- An LLM can turn a query into a rule with the right **structure**, but the rule's **numbers** need calibration.
- Learning-to-rank on **unlabeled in-domain videos** calibrates these numbers without labels from the target data.
- Calibrated rules improve four off-the-shelf retrievers and, together with dense and sparse retrieval, give the **best results** on three driving benchmarks.
- We release human-annotated **motion-event labels for DrivingDojo**.

More Details

Project Page

gyeow.github.io/strive-d

Released Annotations

github.com/nec-labs-ma/strive-d-data-annotation

Supported by NSF grants CMMI-2326309 and CNS-2312395.