MOSEL: Inference Serving Using Dynamic Modality Selection

The University of Texas at Austin
Department of Computer Science
College of Natural Sciences

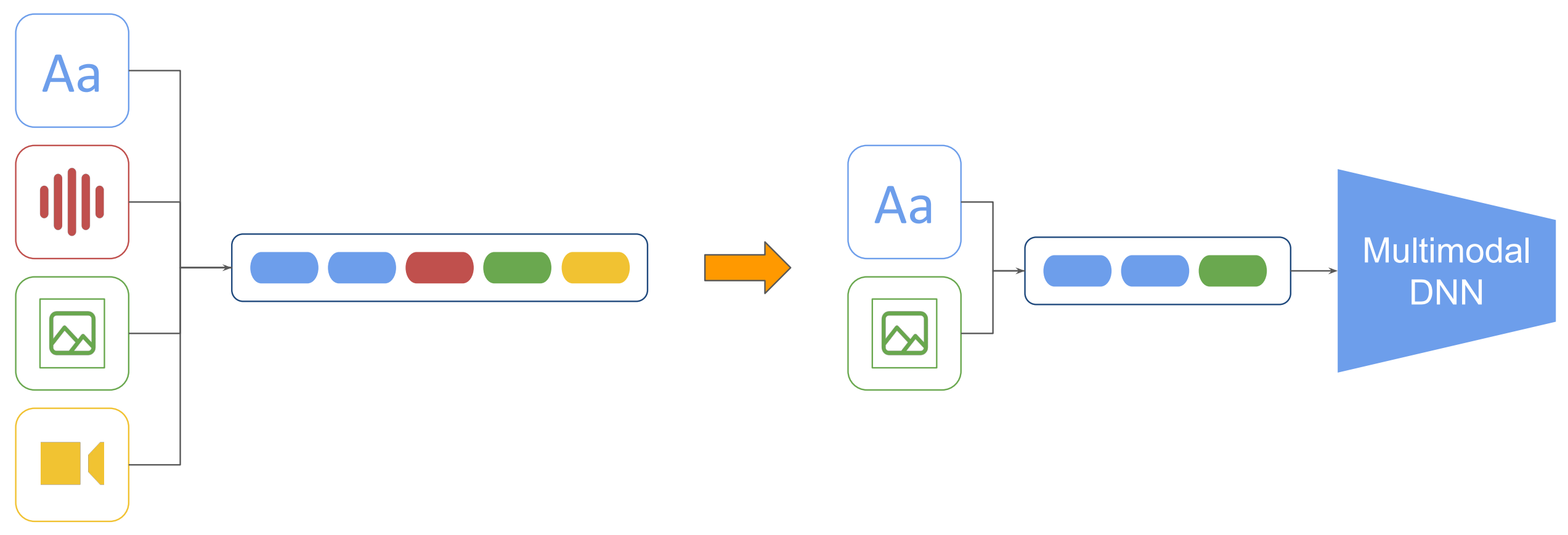
Bodun Hu, Le Xu, Jeongyoon Moon, Neeraja Yadwadkar, Aditya Akella
Department of Computer Science, College of Natural Sciences, The University of Texas at Austin

Summary

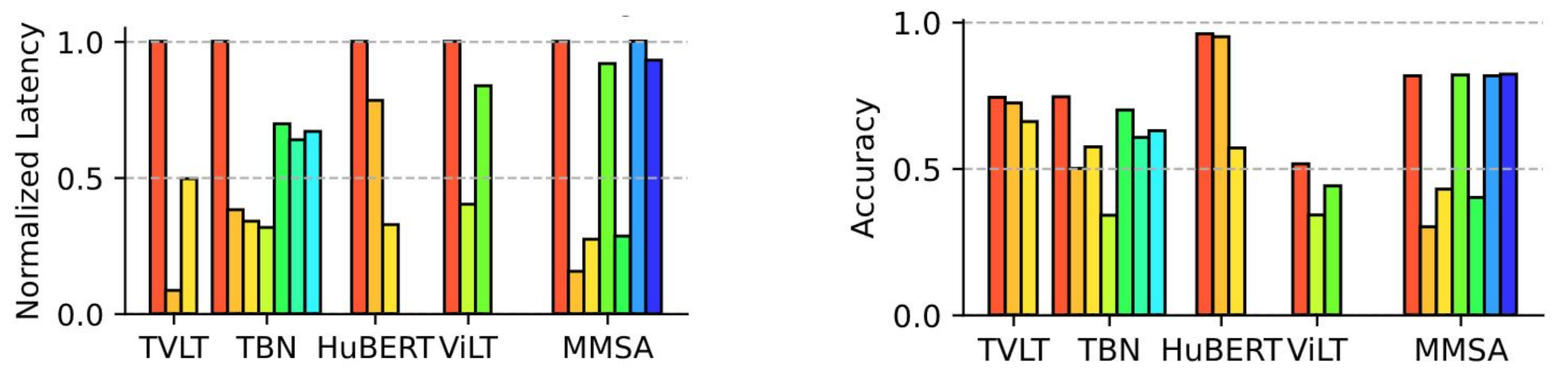
We proposed MOSEL, a framework that automatically select modalities for inference.

- **Based on resource availability, as well as user-defined latency and accuracy requirements.**
- **Improves system throughput by up to 3.6✕ with accuracy guarantees.**

Motivation

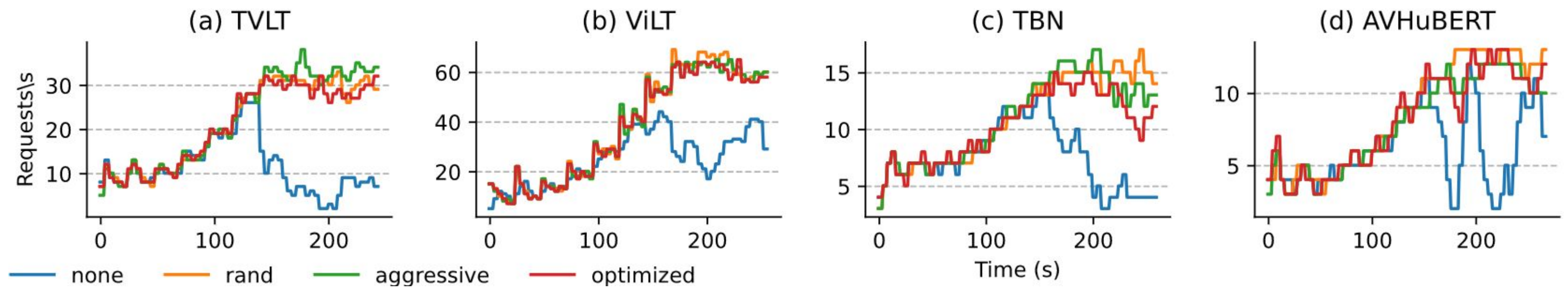


Insight: Dropping modalities can reduce model input size, thus reducing latency.

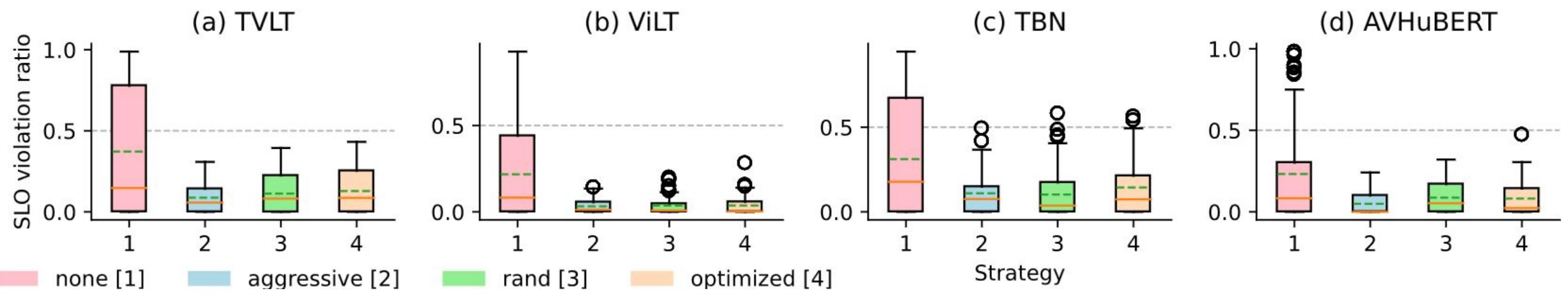


Observation: diverse accuracy and latency trade-offs using different modalities.
Problem: how to navigate the accuracy-latency trade-off space?

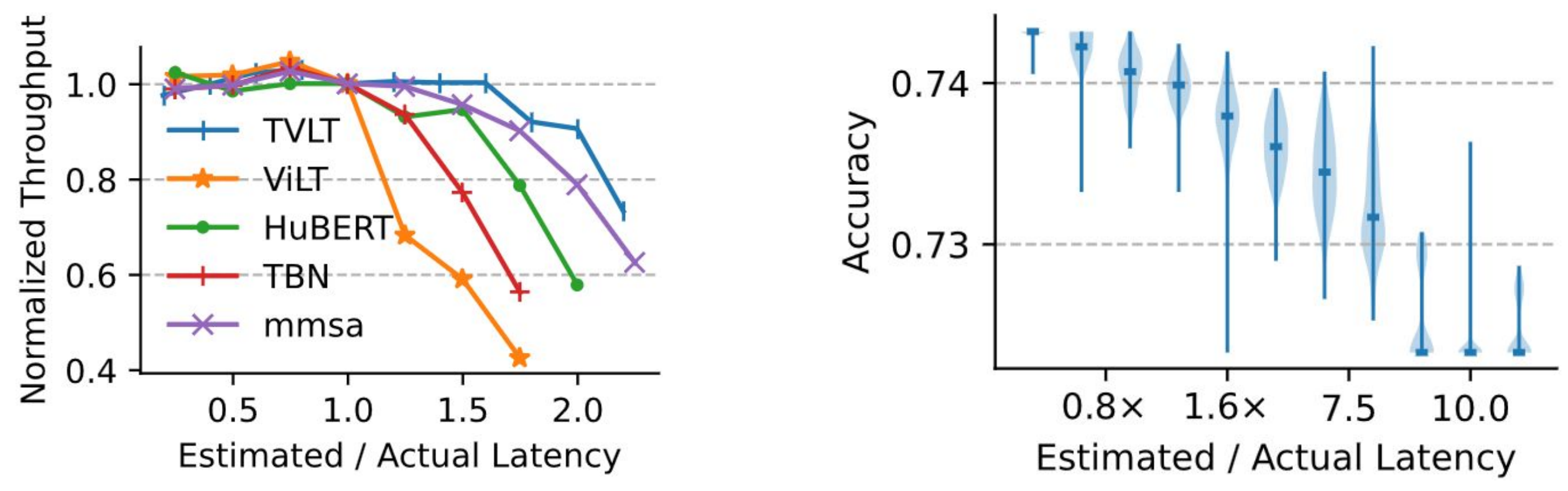
Results



Improved throughput vs modality-agnostic approach on trace data from production environment.

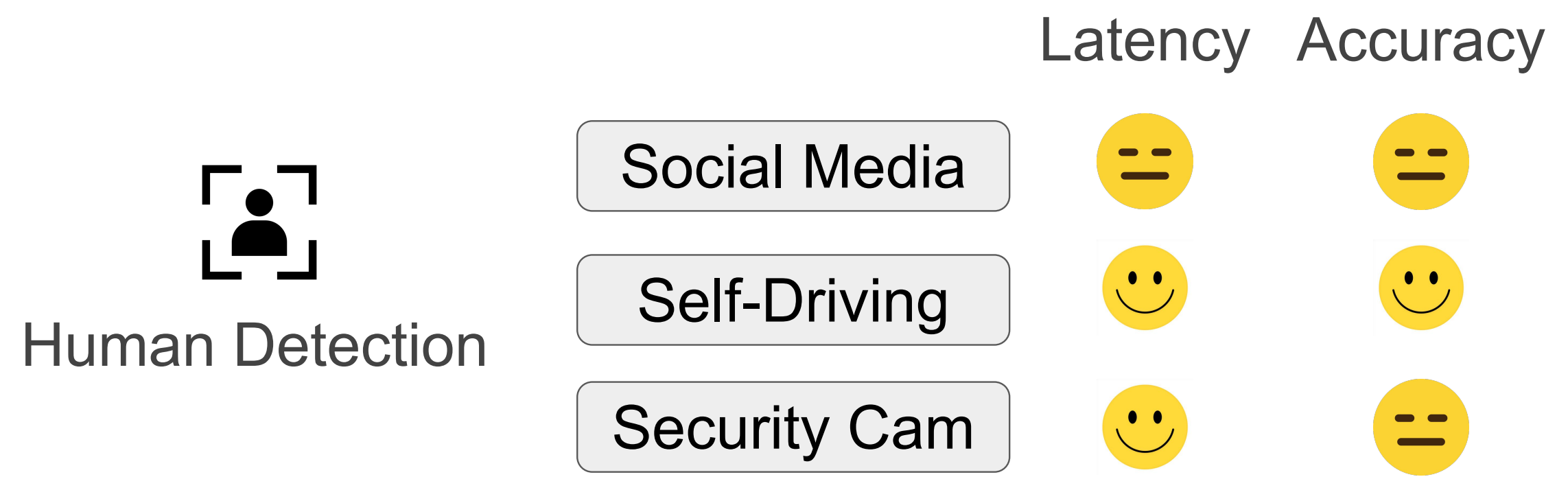


Reduced SLO violation ratio for different types of multimodal models.

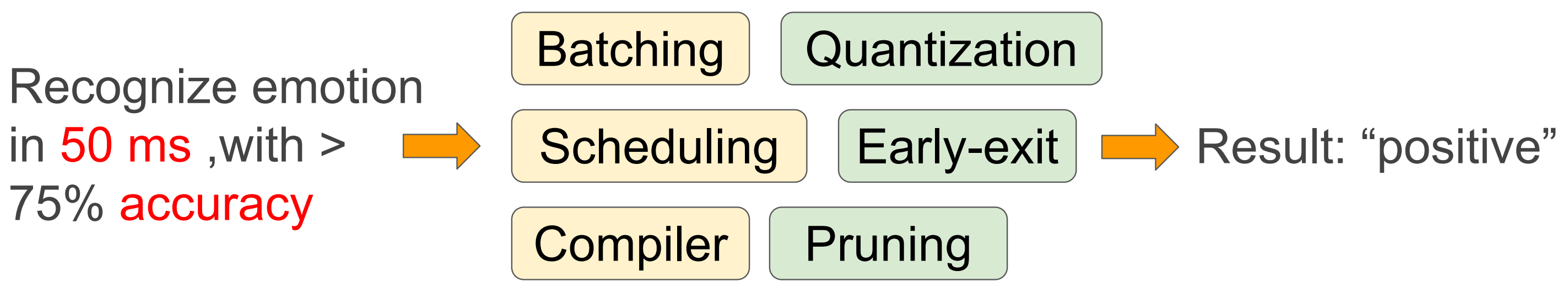


The modality selection plan relies on offline latency profiling. MOSEL tolerates errors in these profiles without compromising system throughput and accuracy. Significant errors are rare since DNN inference is deterministic with predictable GPU latency.

Introduction



Requirement: Different applications required different **accuracy** and **latency** Service level Objectives (SLOs).


Recognize emotion in 50 ms ,with > 75% accuracy

Result: “positive”

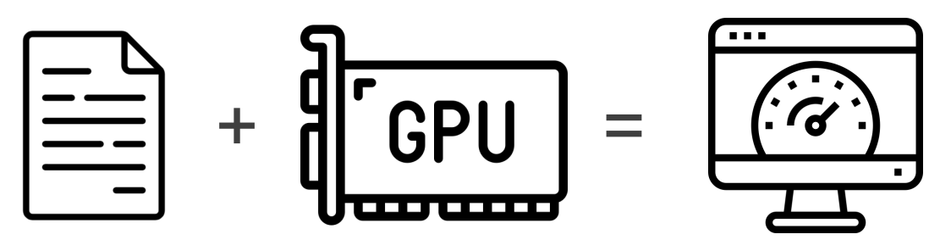
Problem: (1) big search space. (2) rely on expensive hardware. (3) complex system design. (4) required multiple model replicas.

Approach

Goal: Choose modalities to achieve best accuracy without violating latency requirement.

Step 1: Profile accuracy and latency of different modality(ies).

- Accuracy: Derived from datasets, independent of batch size, dependent on modality choices.
- Latency: Measured on real hardware, dependent by batch size, dependent on modality choices.

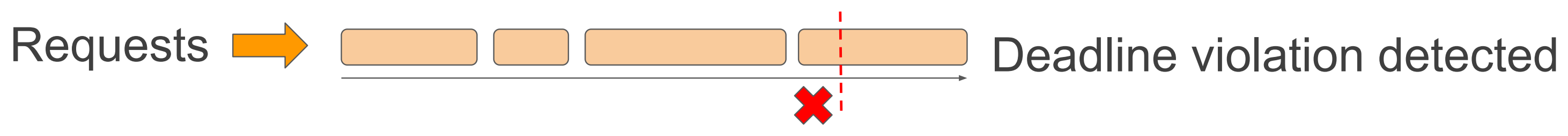


Modality accuracy and execution latency over different batch sizes.

Step 2: Generate **optimal** modality selection plans using profiled data.

- Since accuracy SLO is unknown at inference time, create **optimal** plans for **all achievable** accuracy SLOs. An **optimal** plan meets a given accuracy SLOs with the lowest latency.
- These optimal plans can be queries by (batch, accuracy).
- Generated using Integer nonlinear programming (INLP).

Step 3: Dynamically select modalities for requests to **maximize** accuracy while adhering to latency constraints. This step updates each request with a new accuracy SLO.

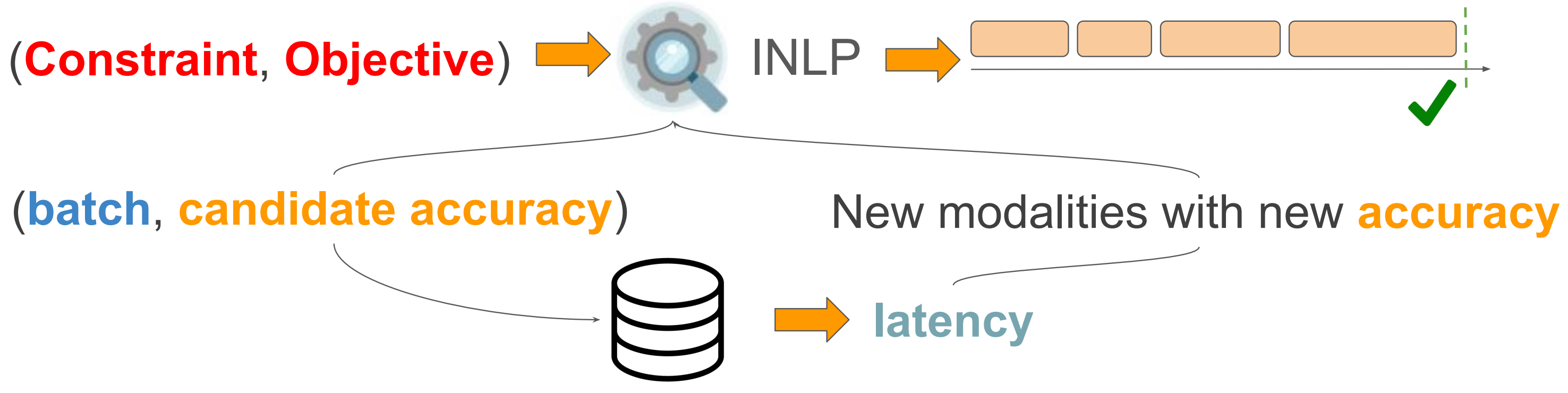


Constraint: Total latency of all requests must be under T.

$$\sum_{s \in S} l(s) \leq T$$

Objective: use available resource (T**) to maximize accuracy.**

$$\sum_{s,j \in S,J} acc(s) \cdot |j|$$



When a deadline violation occurs, MOSEL adjusts to meet latency constraints and maximize accuracy. It queries the database of optimal modality plans to verify if the candidate plan's latency adheres to the constraint.