Scalable Video Processing Using Spark

Siddhant Garg & Sridhama Prakhya

COMPSCI 532: Systems for Data Science

University of Massachusetts Amherst

Content

- **Motivation and Introduction**
- **Components**
  - Spark
  - CLIP
  - MongoDB
- **Experiments and Results**
- **Conclusion and Future Work**

Motivation: Raw video datasets

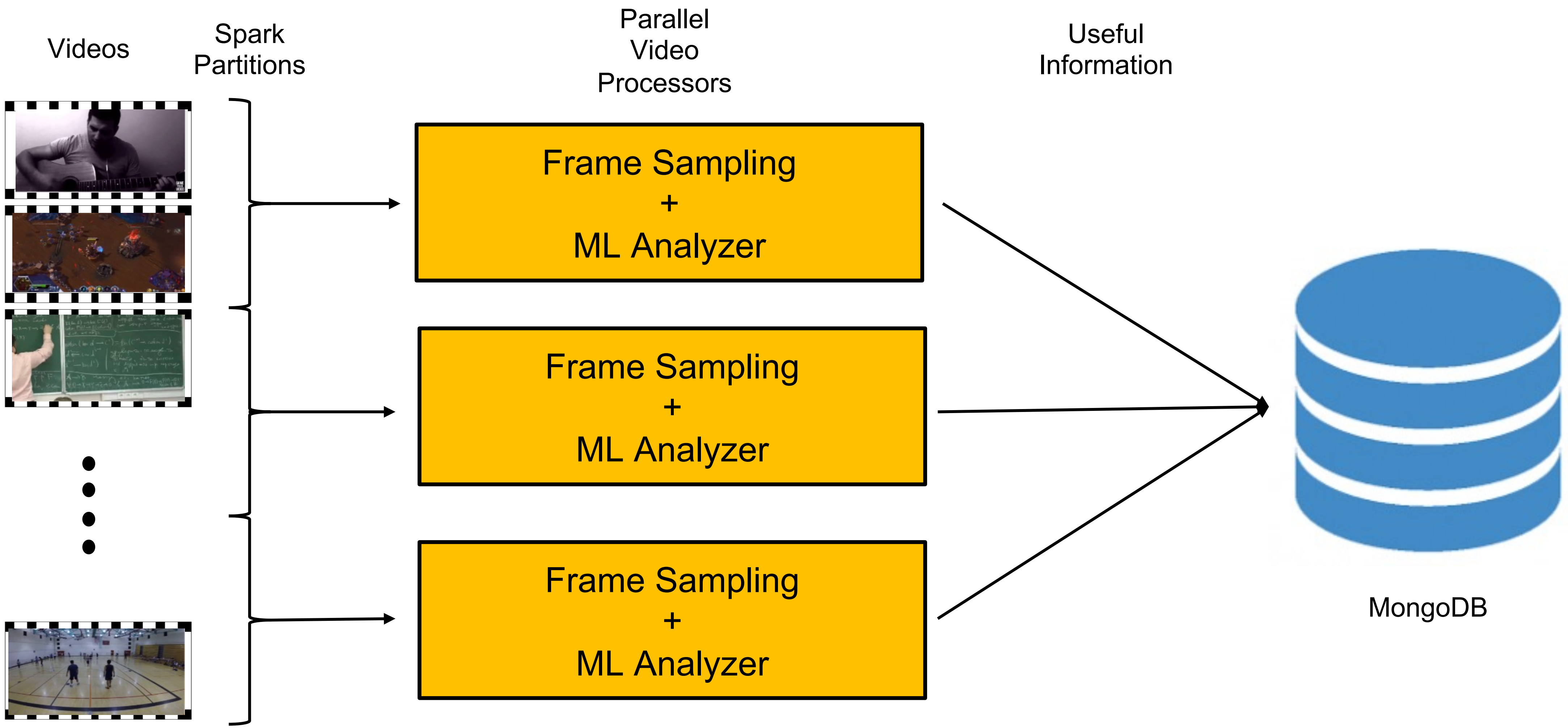
Motivation

- **Video Processing**
  - Frame sampling
    - 9K frames for 5-minute video @ 30 FPS
  - Machine Learning inference
    - Batch inference
  - Takes too long for large datasets
    - 1 month for 500K videos on a single machine
- **Analysis**
  - Retrieve videos from specific classes
  - Retrieve similar videos based on shared features

Introduction

- **Parallelize video processing using Spark**
  - Automatically partition data across machines
  - Run feature extraction on each machine in parallel
  - Machine Learning model
    - CLIP model
    - ImageNet classes (1K)
    - Kinetics (700 action classes)
- **MongoDB database**
  - Storing extracted tags for each video
  - Efficient and simple for analyzing extracted information

Overview


Videos
Spark Partitions
Parallel Video Processors
Useful Information
Frame Sampling + ML Analyzer
Frame Sampling + ML Analyzer
Frame Sampling + ML Analyzer
MongoDB

Database storage

- MongoDB
  - NoSQL database
    - Horizontal scalability
    - Fast retrieval for semi-structured data (tag arrays)

- SparkSQL for insertion and querying
  - Simple retrieval query based on single tag
  - Conditional retrieval using multiple tags

Frame Sampling

+

ML Analyzer

Video Processing

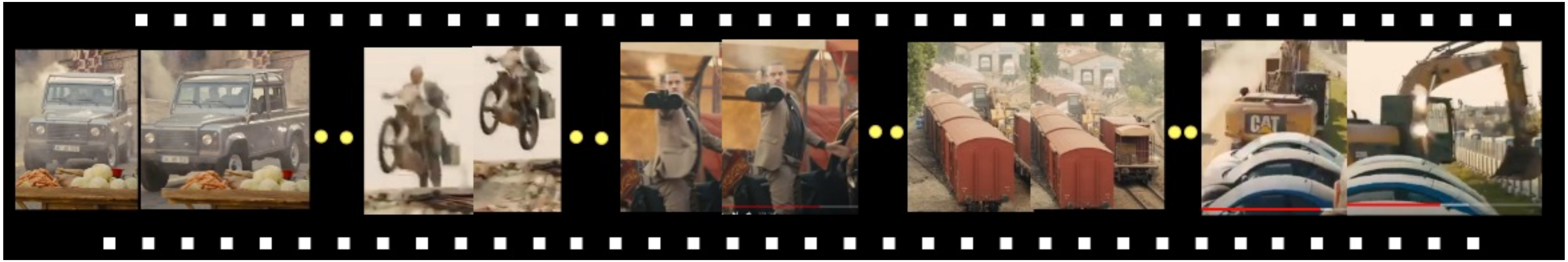
Duration: 4 mins
Frame rate: 30 FPS

TOTAL FRAMES: 7200

Source: https://www.youtube.com/watch?v=u3dqiq0uFdg

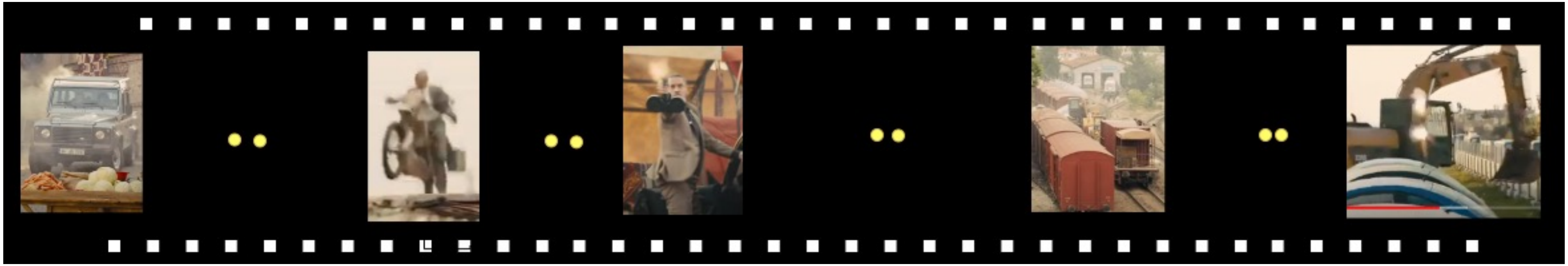
Video Processing

Total Frames: 7200



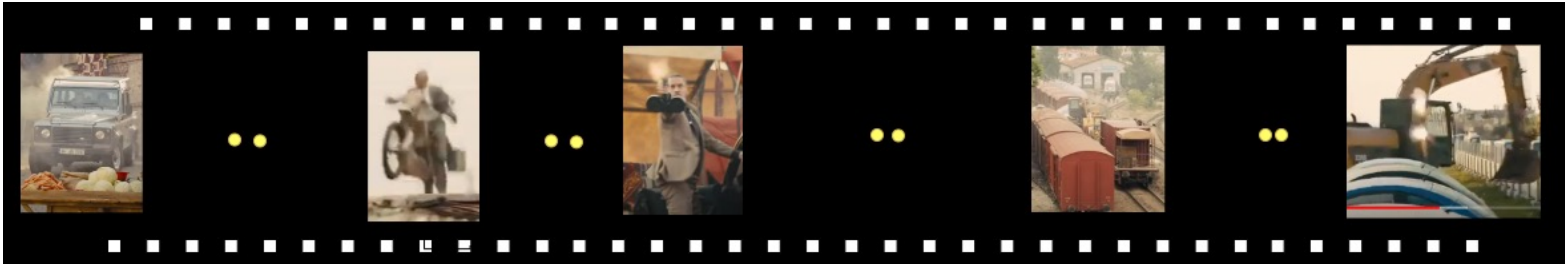
Frame Sampling @ 1 FPS

Sampled Frames: 240



Video Processing

Sample Frames: 240



ML Analyzer

[jeep, parkour, shooting off fireworks, freight car, bulldozing]

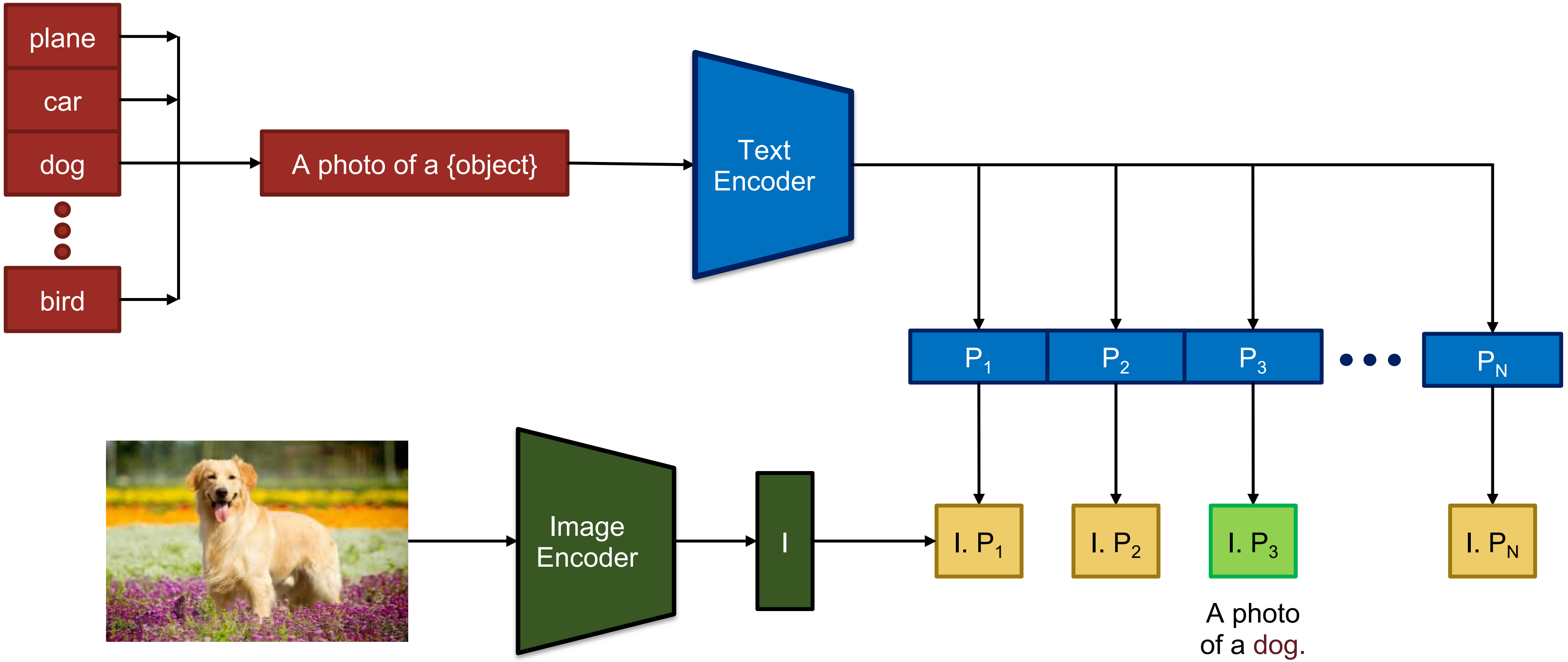
Frame Sampling
+
ML Analyzer

Machine Learning analyzer

- CLIP model
    - Text Transformer encoder
    - Image Transformer encoder
    - Trained using natural language supervision
    - Support zero-shot predictions

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International Conference on Machine Learning*. PMLR, 2021.

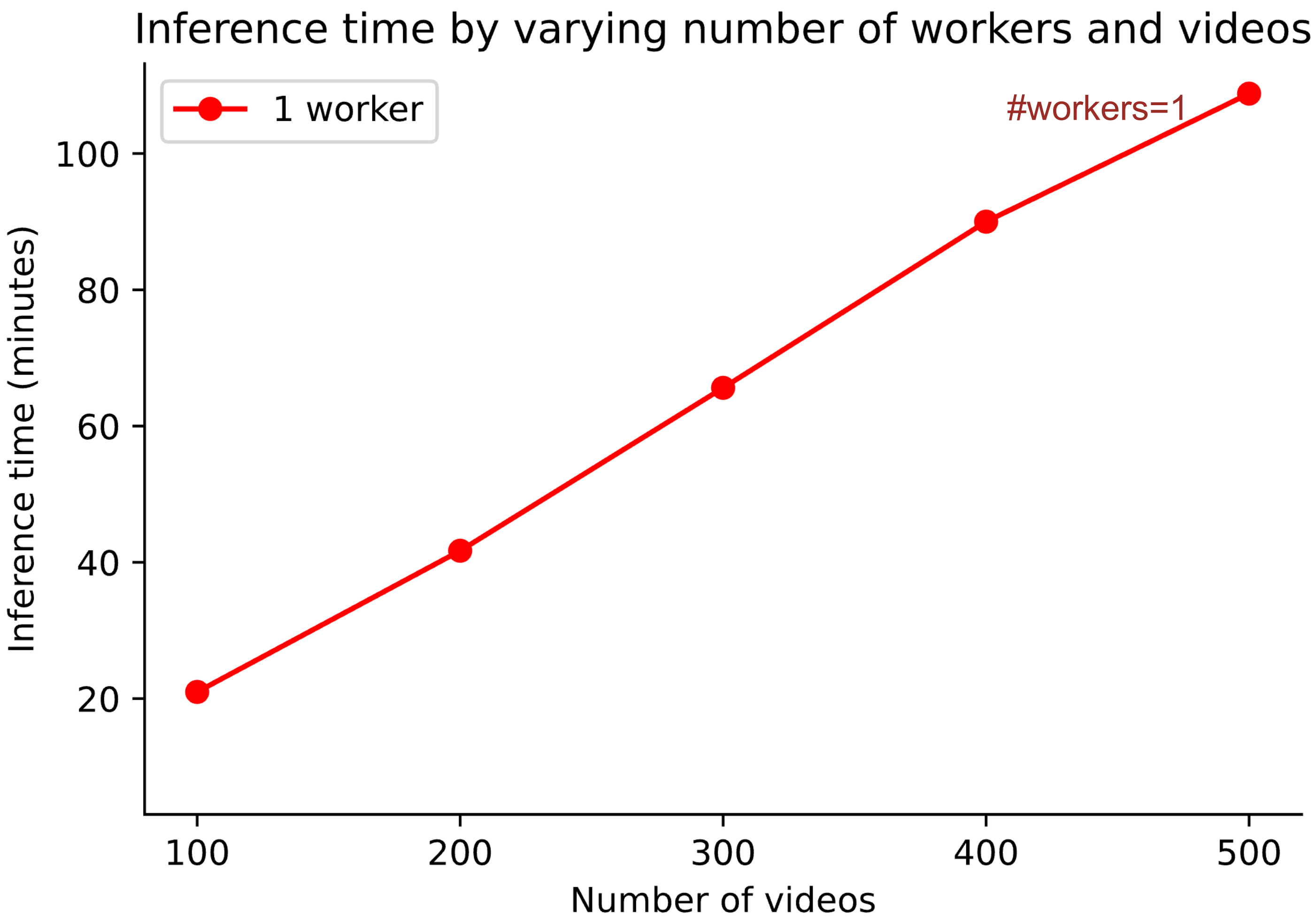
CLIP Zero-Shot Prediction



Experiments and Results

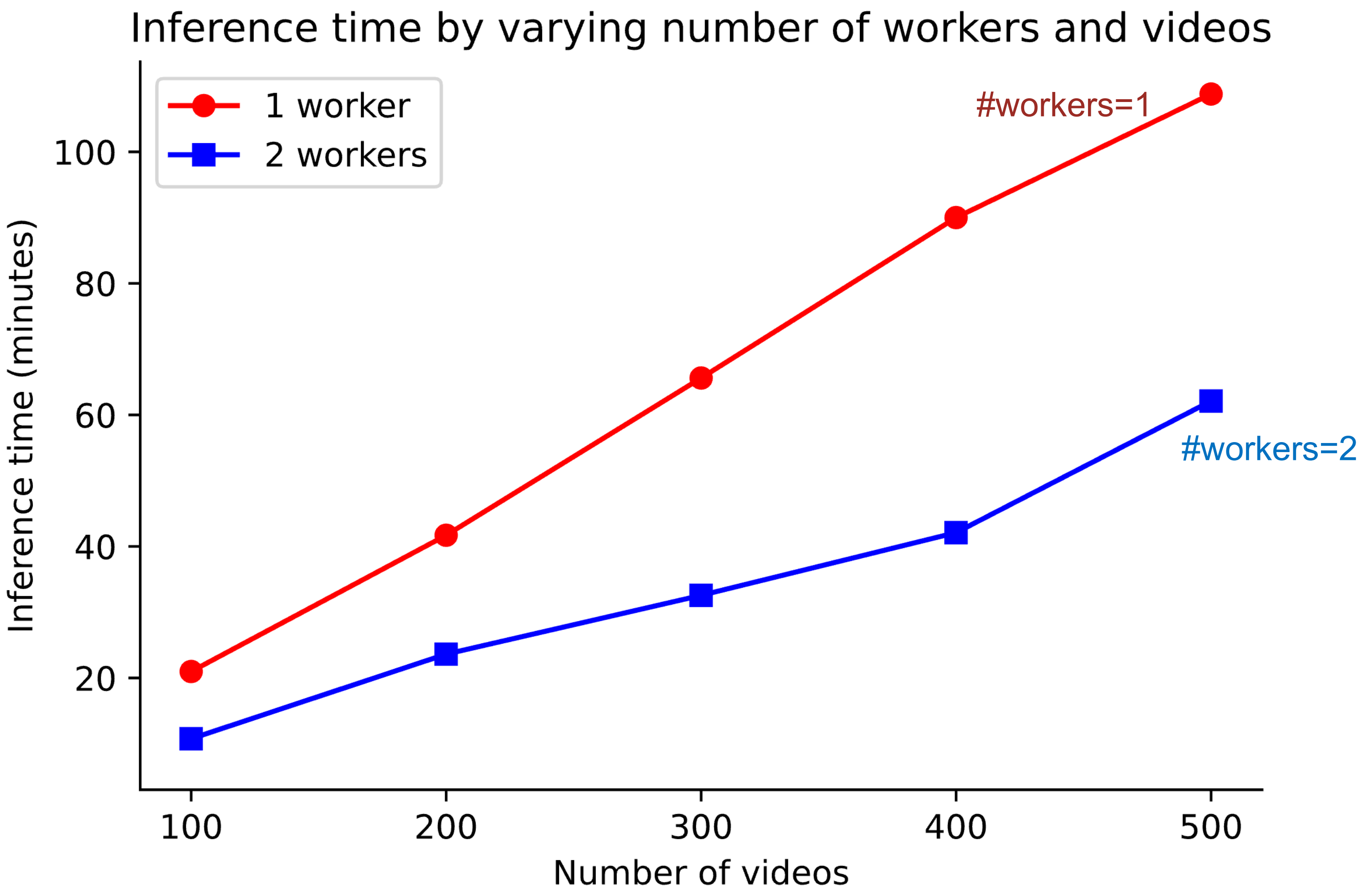
University of Massachusetts Amherst

Inference Durations



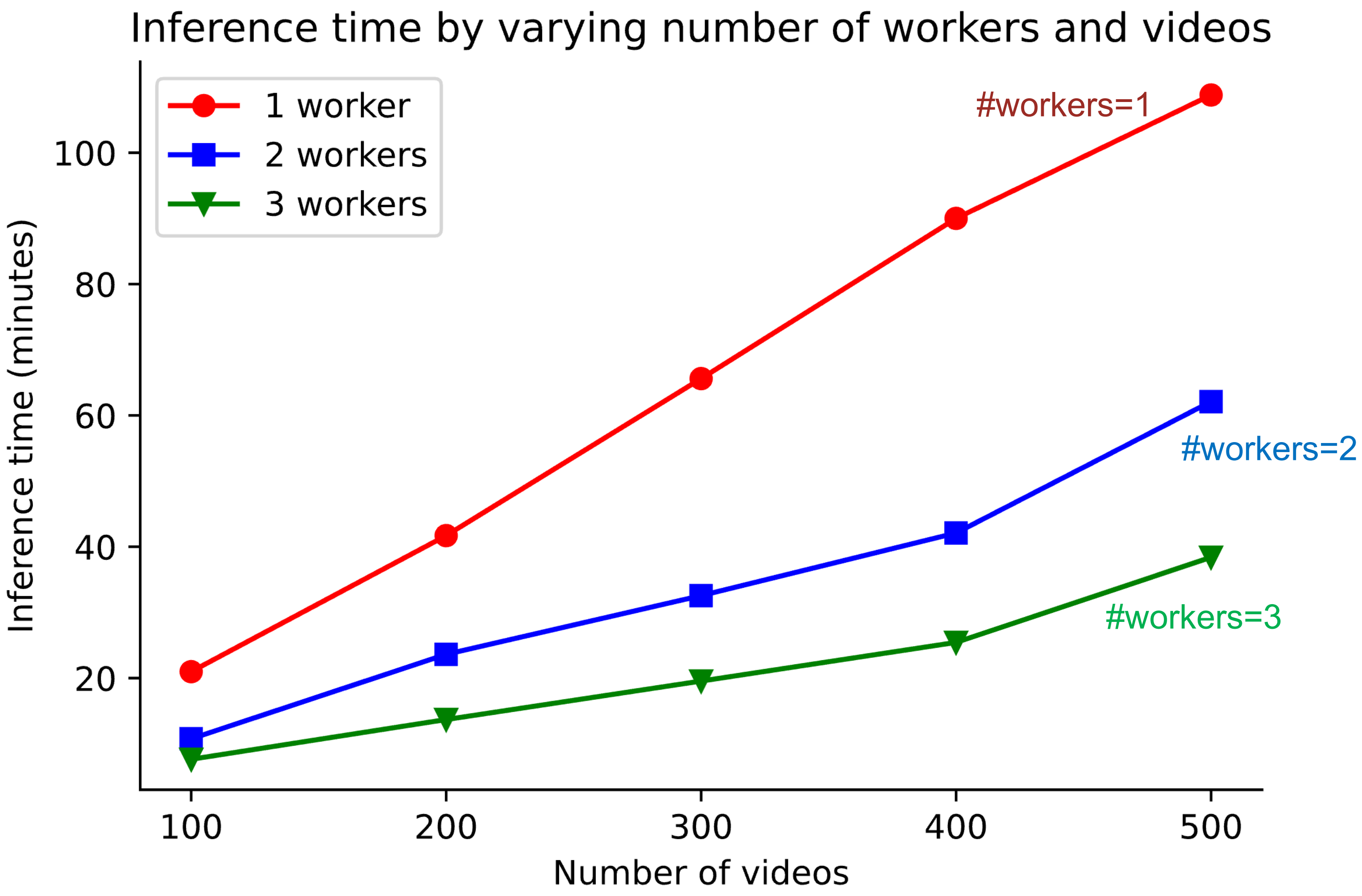
GPU : 1 Tesla T4, 15GB

Inference Durations



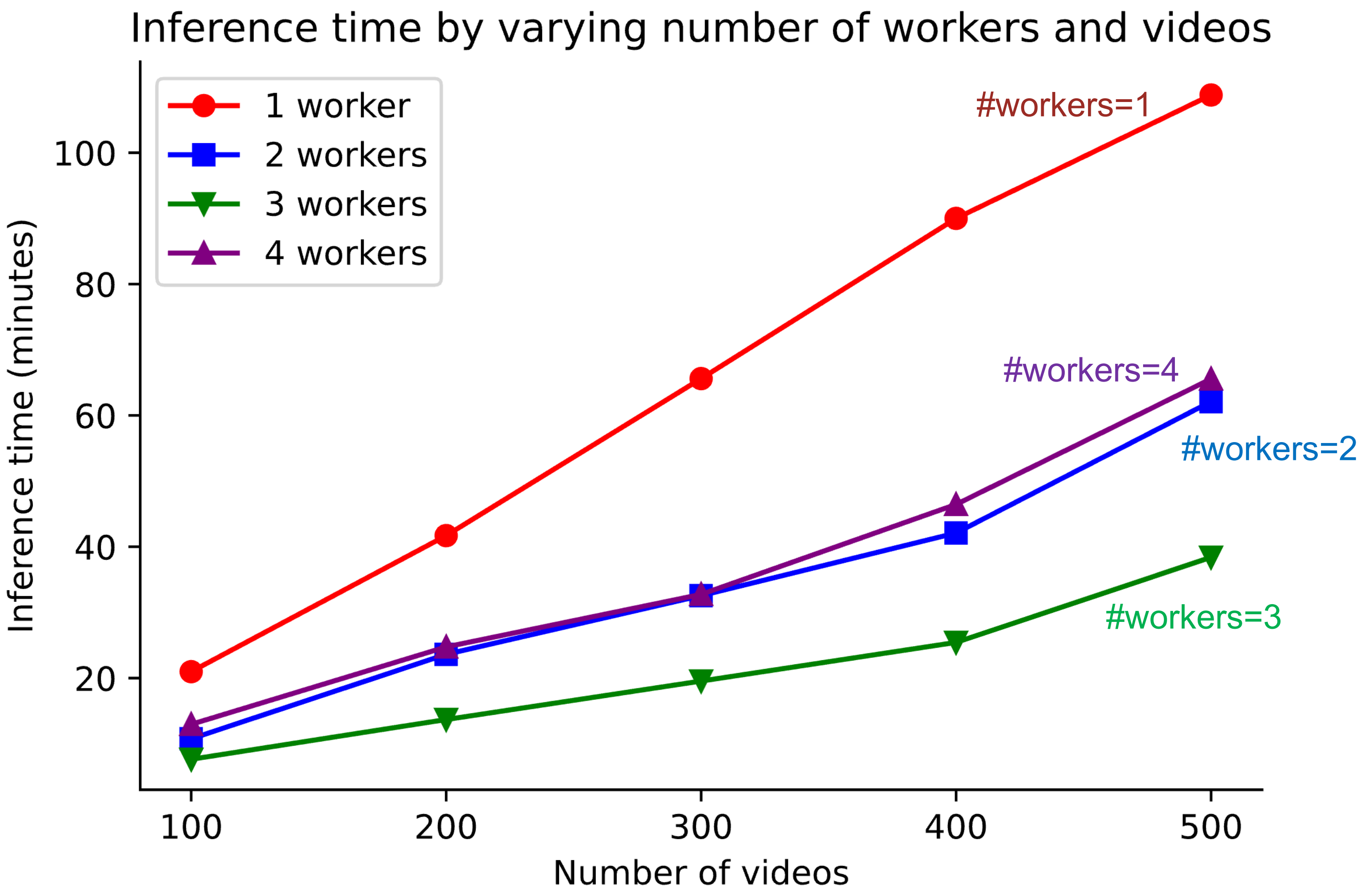
GPU : 1 Tesla T4, 15GB

Inference Durations



GPU : 1 Tesla T4, 15GB

Inference Durations



GPU : 1 Tesla T4, 15GB

Query Results for ‘playing cello’ (video-1)

Cello/Playing cello

Drums/Playing Drums

Violin/Playing violin

Source: https://www.youtube.com/watch?v=j7MZfgP3j7k

Query Results for ‘playing cello’ (video-2)

Cello/Playing cello

Cello/Playing cello

Bartending

Source: https://www.youtube.com/watch?v=Fzlh69nLx5Y


Thank you
University of Massachusetts Amherst
BE REVOLUTIONARY™