ARTIFACT EVALUATED usenix ASSOCIATION AVAILABLE

ARTIFACT EVALUATED usenix ASSOCIATION FUNCTIONAL

ARTIFACT EVALUATED usenix ASSOCIATION REPRODUCED

CoVA: Exploiting Compressed-Domain Analysis to Accelerate Video Analytics

Jinwoo Hwang, Minsu Kim, Daeun Kim
Seungho Nam, Yoonsung Kim, Dohee Kim
Hardik Sharma*, Jongse Park

KAIST, *Google

CASYS | KAIST Computer Architecture & System Lab

Growing Video Data

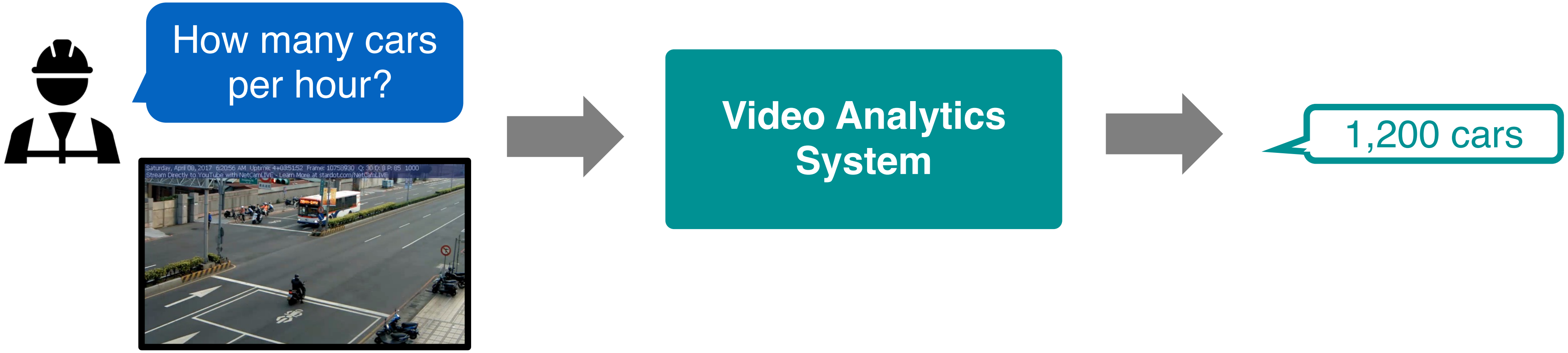
Video data makes up **82% of global IP traffic*** as of 2022, and is **growing**

* CISCO Annual Internet Report

Video Analytics

Video Analytic System analyzes video to extract high-level information and answers **user queries**

Example



Using Object Detector for Video Analytics

Frame 1 **Frame 2** **Frame 3**

When does a ***car*** appear?

Answer: Frame 3!

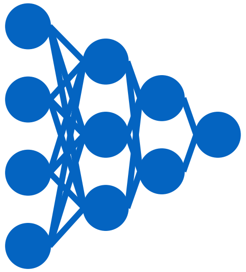
Video Analytic System

Object Detector

Frame #	Class	X, Y, W, H
2	Person	(140, 550, 130, 100)
3	Person	(870, 570, 140, 100)
3	**Car**	**(150, 410, 24, 64)**

Challenge and Prior Approaches

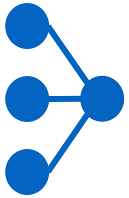
Challenge



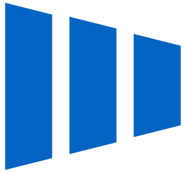
DNN-based object detector requires **heavy computation**

e.g., YOLO take 11 hours to process two weeks long video

Prior Approaches [VLDB'18, ICDE'20, VLDB'20]



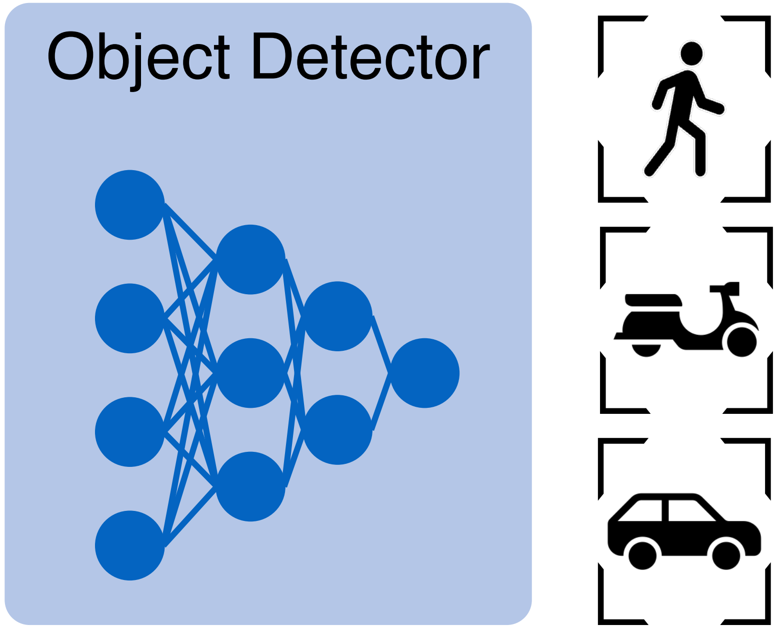
Simple neural networks ***specialized*** for the user query



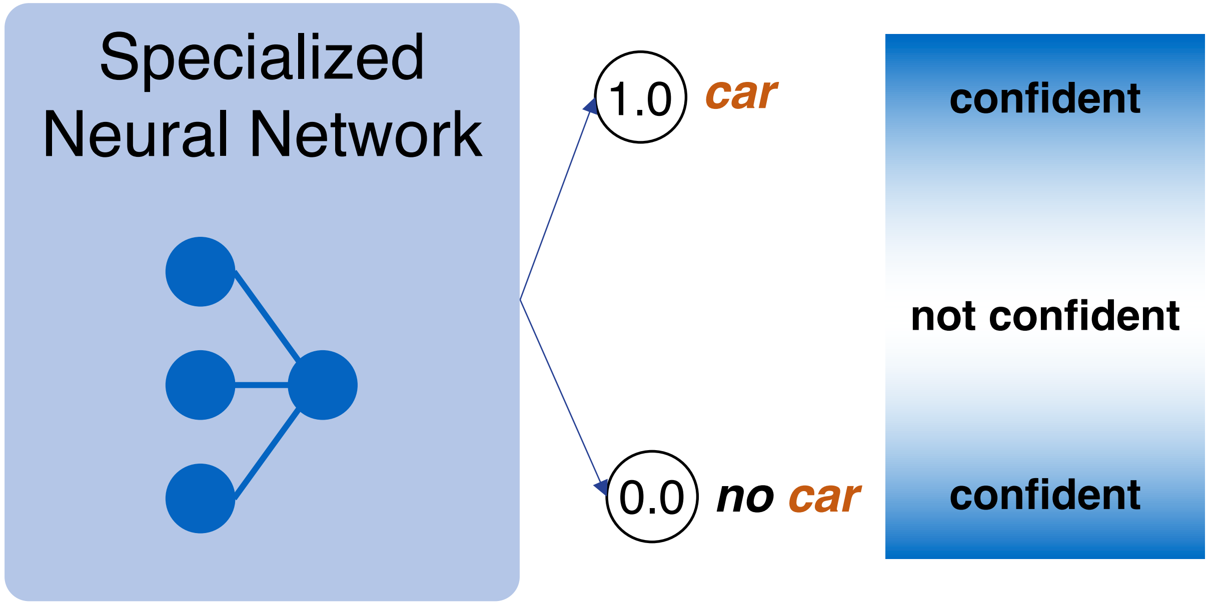
Cascade architecture constitute a pipeline of classifiers that trades accuracy and performance

Prior Approach: Specialized Neural Network

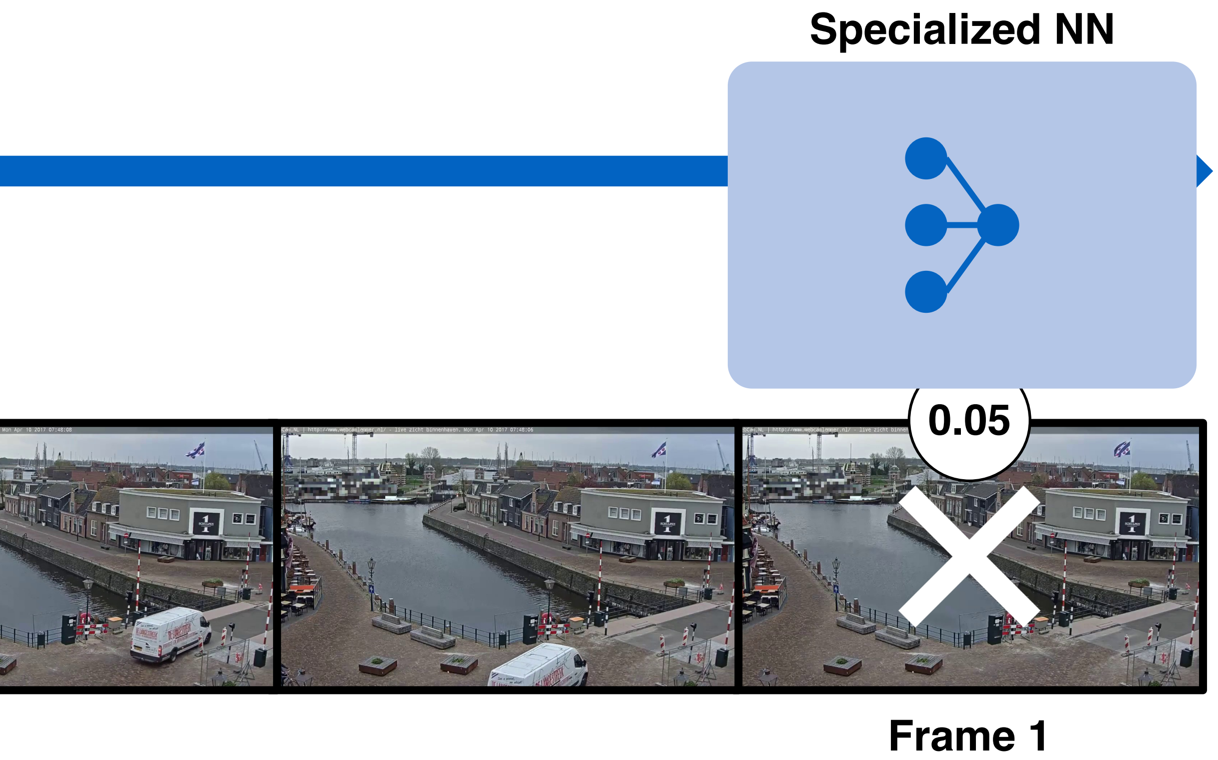
Complex Task
Determine location and class of ***every object***



Simple Task
Determine if there is ***any car*** or not

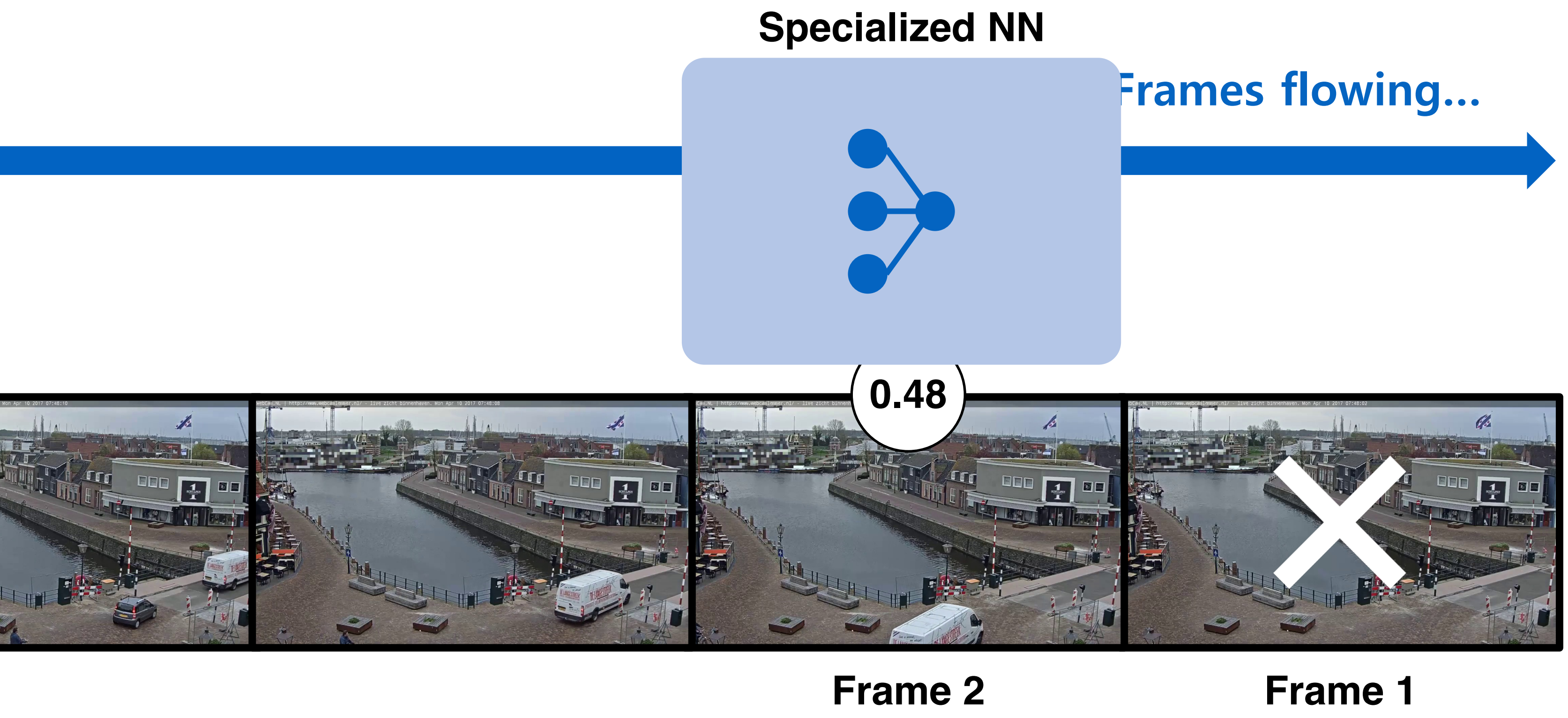


Prior Approach: Cascade Architecture



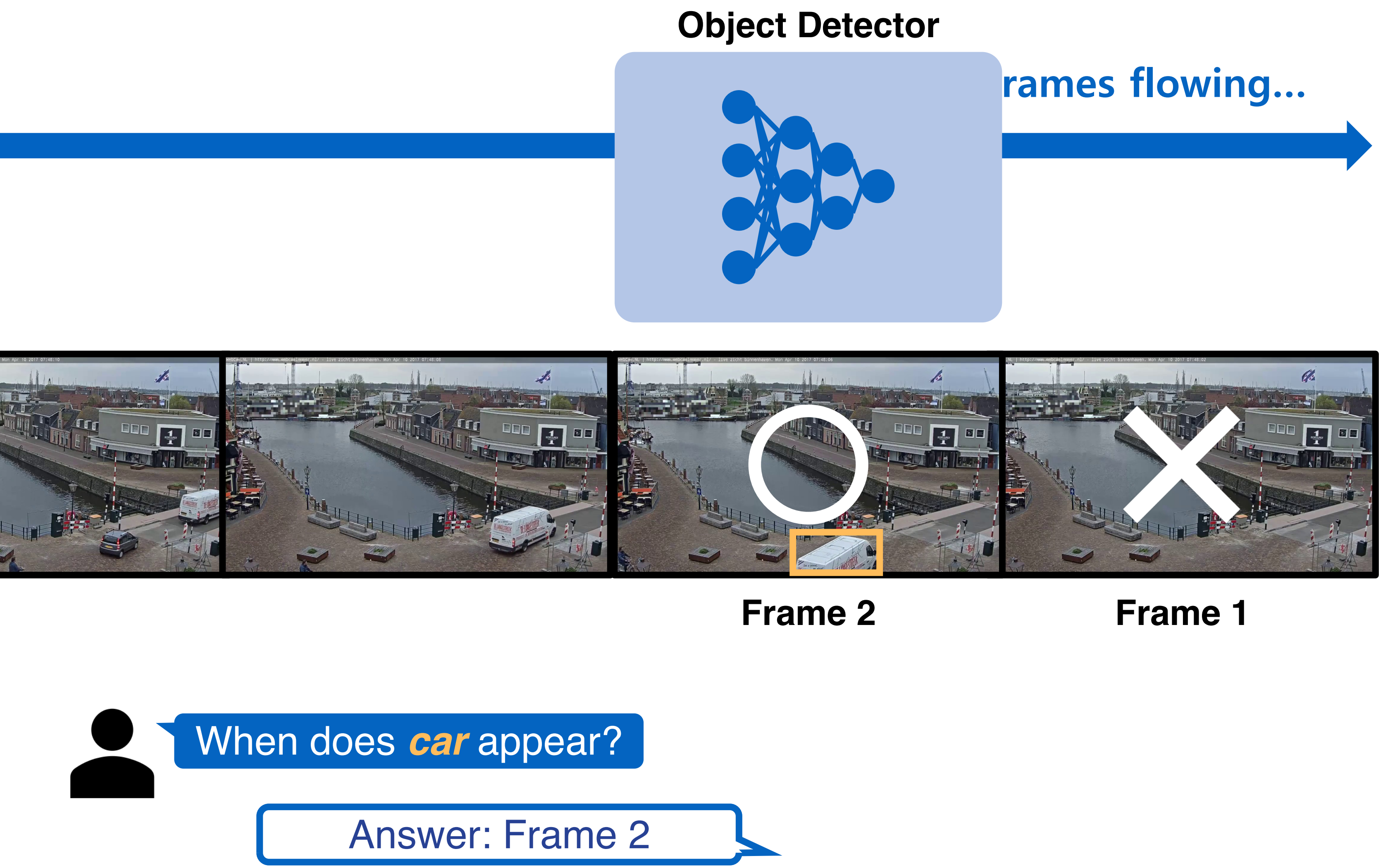
When does *car* appear?

Prior Approach: Cascade Architecture

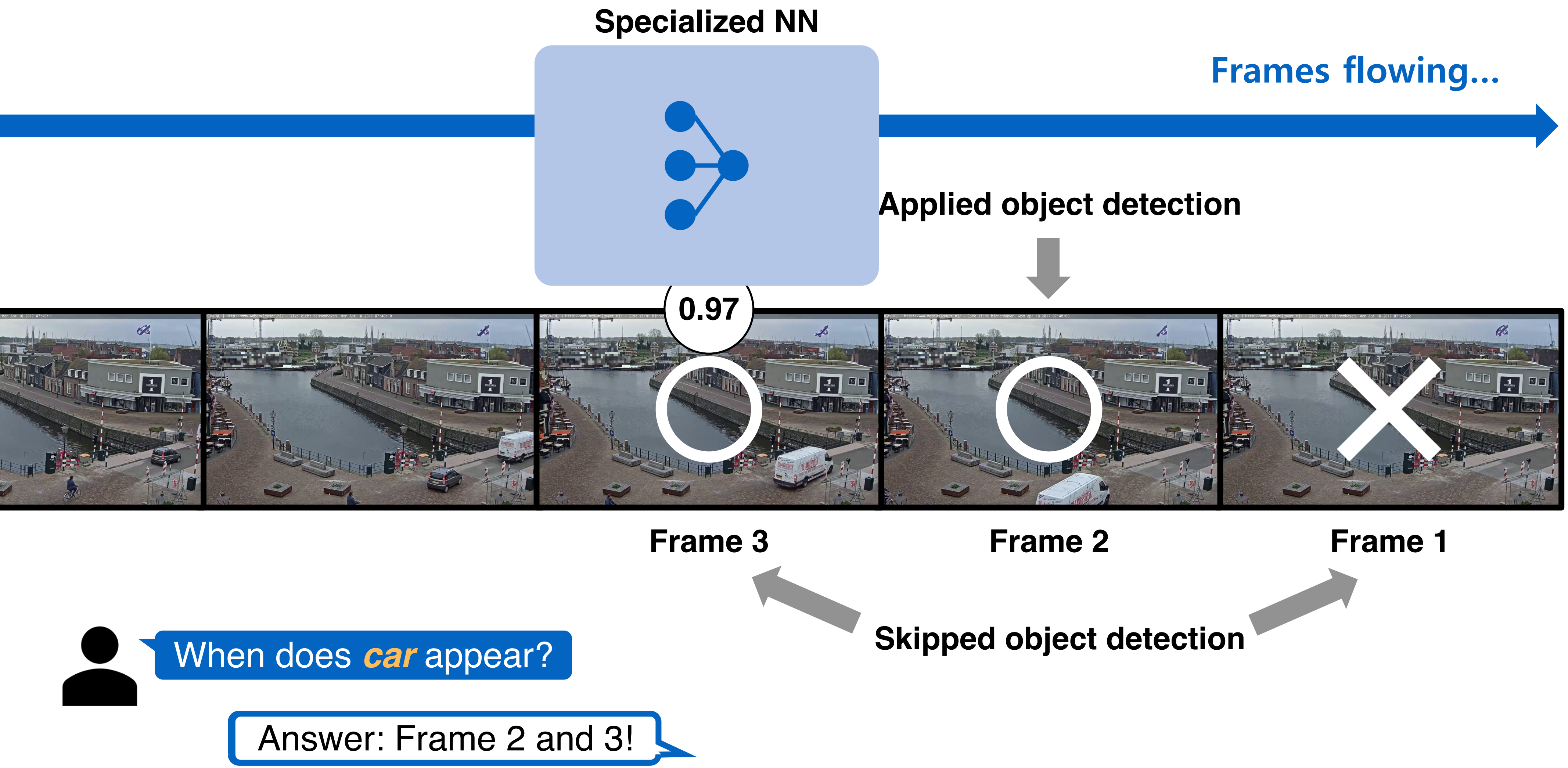


When does *car* appear?

Prior Approach: Cascade Architecture



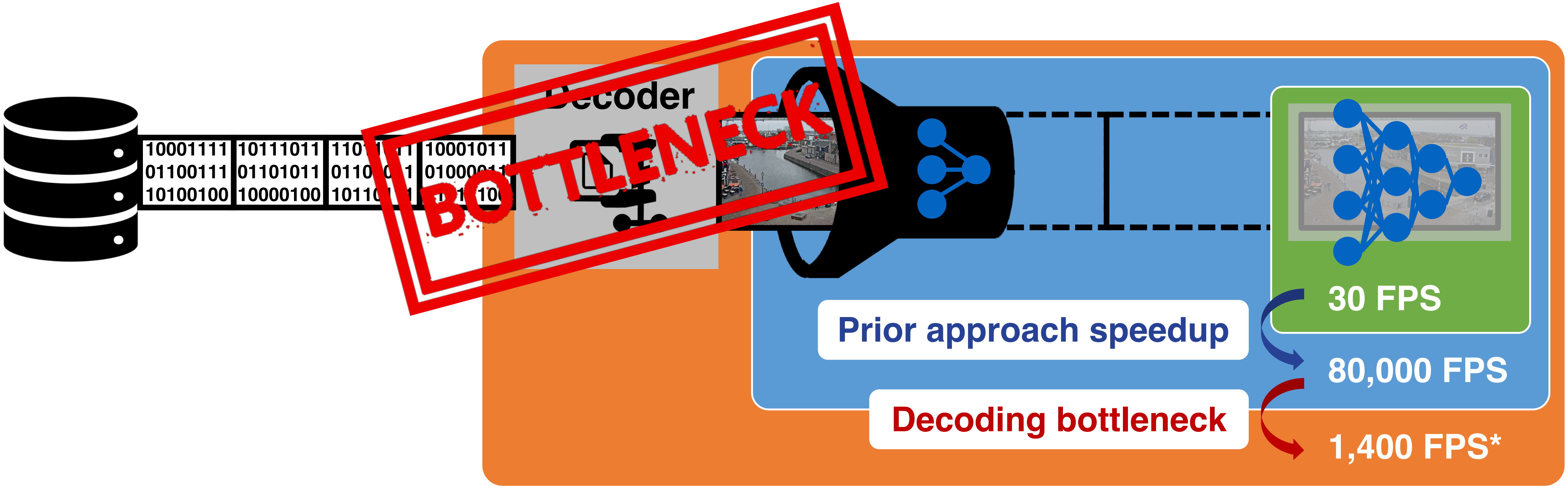
Prior Approach: Cascade Architecture



Two Limitations of Prior Approaches

1. Bottleneck from Decoding

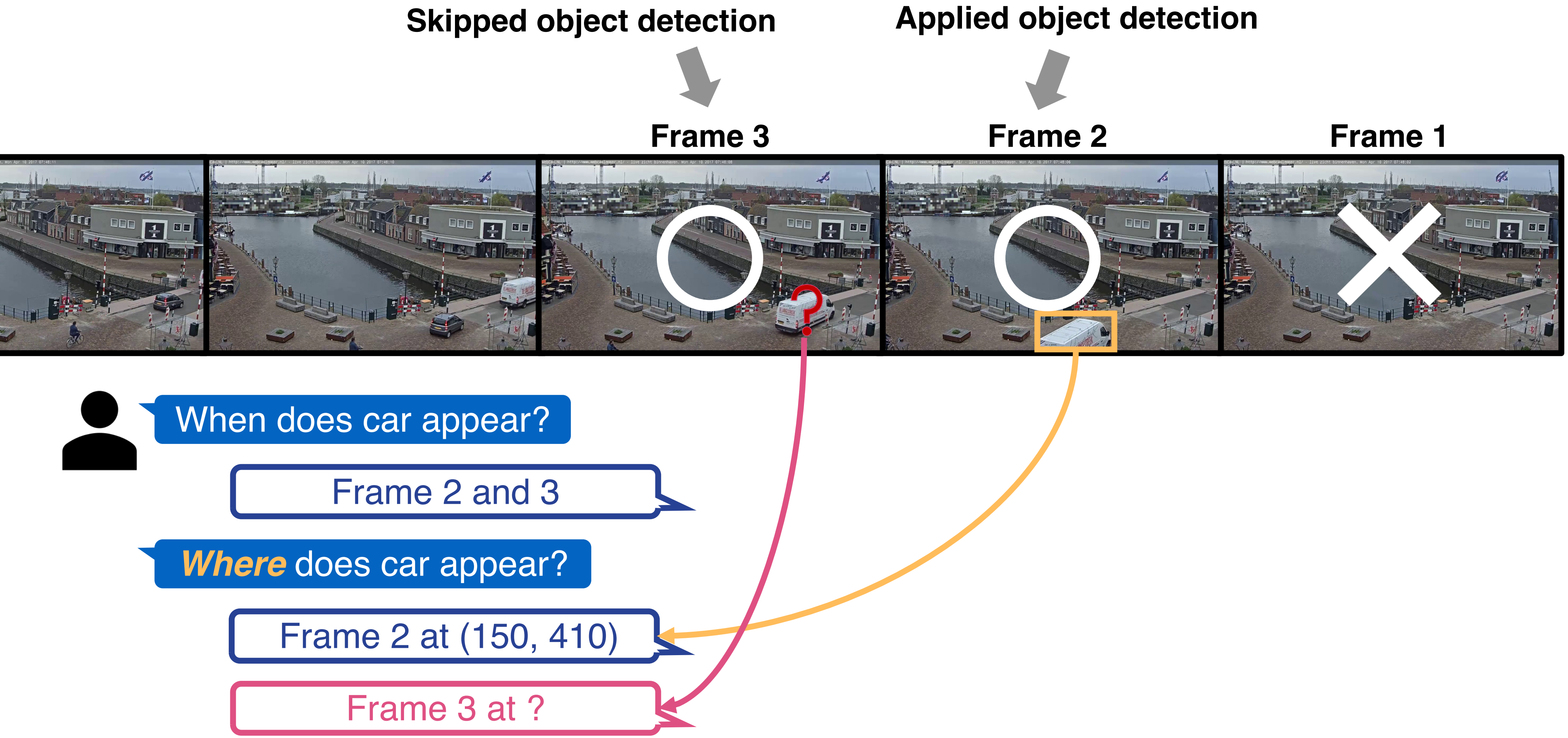
- Prior works ignore a compute-heavy preprocessing stage, **video decoding!**



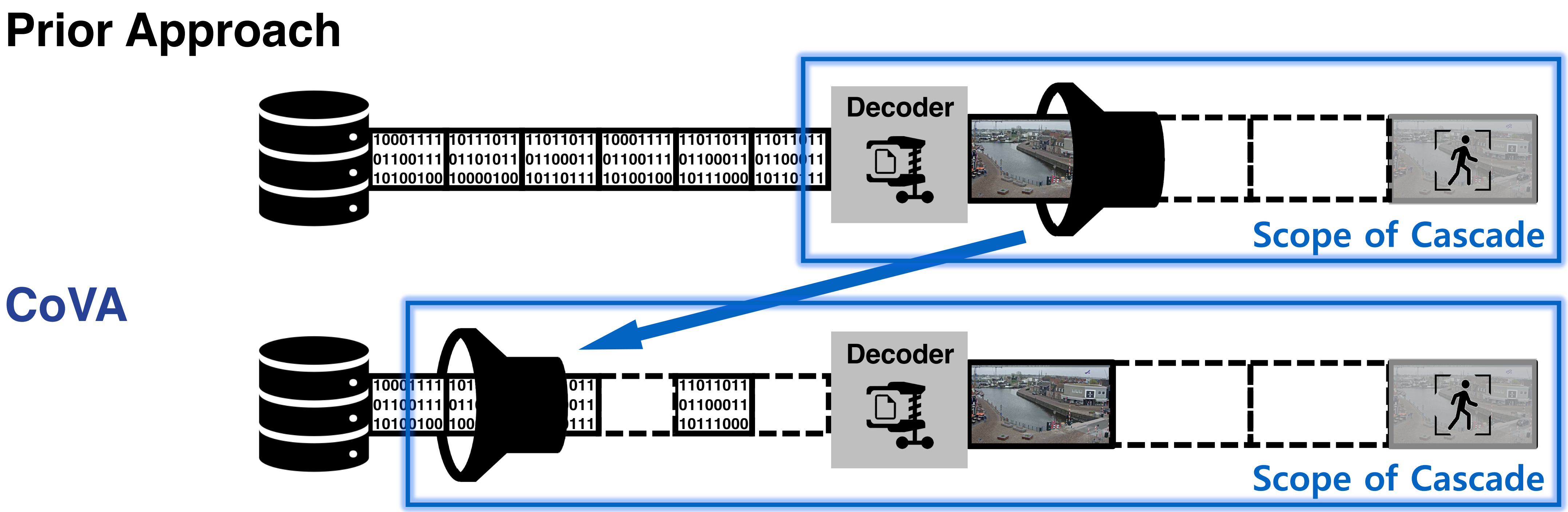
*** 720p video with HW acceleration, NVDEC**

Two Limitations of Prior Approaches

2. Lack of Support for Spatial Query



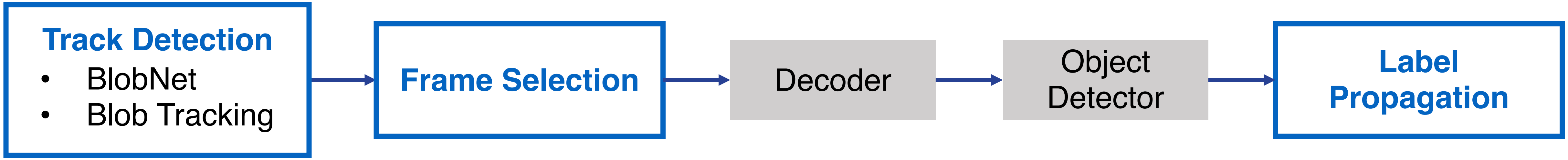
CoVA: Compressed Video Analysis



***Contribution 1*: 4.8× end-to-end speedup by addressing decoding bottleneck**

***Contribution 2*: Spatial query support**

CoVA Overview



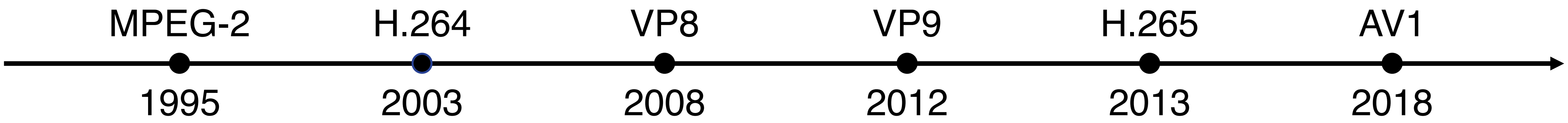
Track Detection

Goal of Track Detection

Goal: **without decoding,** find track of moving objects

How can we find moving objects from compressed video?

How modern video codecs works



Algorithmic commonality: ***Block-based compression***

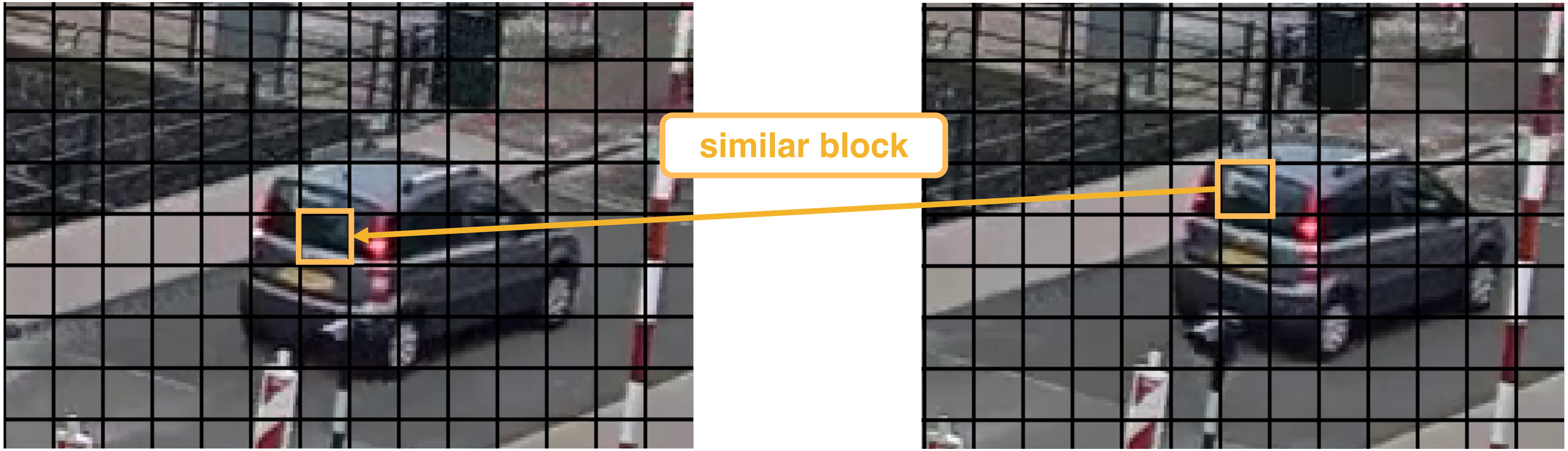
Block-based Compression: Macroblock

Frames are first divided into a grid of *macroblocks*

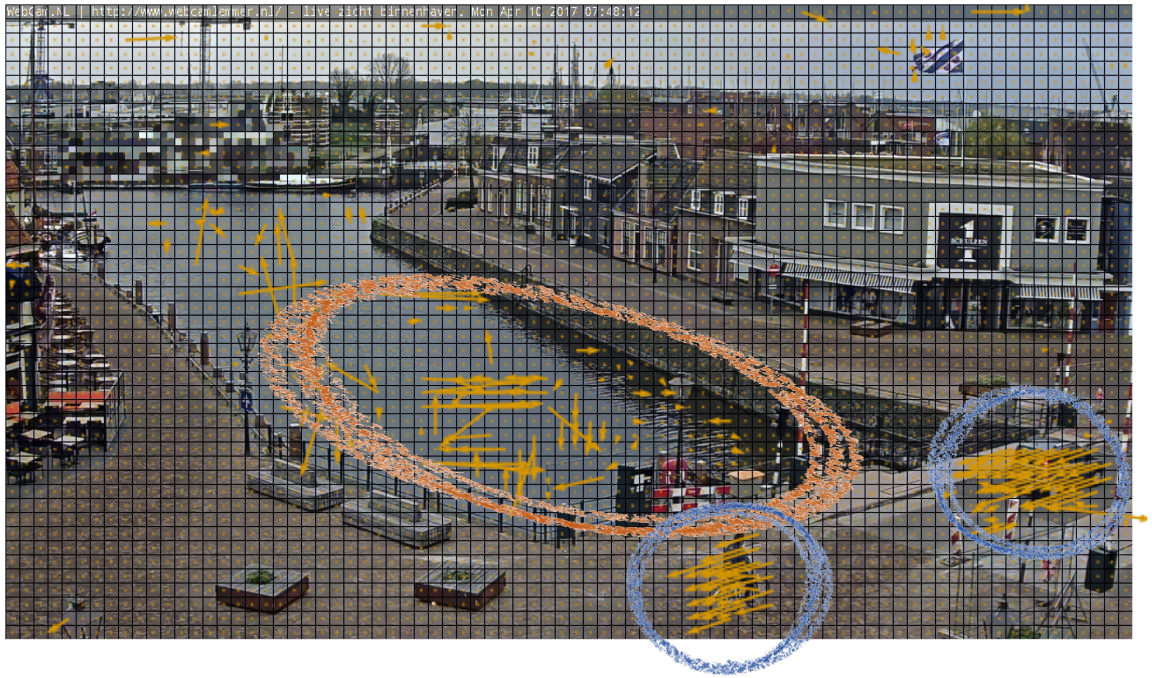
Block-based Compression: Motion Vector

Macroblock is compressed by saving relative position to similar block

Previous frame **Frame to encode**



Challenge in Using Compression Metadata



Challenge: Find moving object from noisy compression metadata

Solution: Neural network based algorithm

BlobNet

Input

- Compression Metadata
  - Motion vector
  - MB type*
  - MB partition*

Embedding Layer

- Additional layer for neural network to **embed compression metadata**

Temporal U-Net

- **Encoder-decoder architecture** for denoising
- Video instance segmentation model architecture running **in pixel domain**

Output

- Training label generated using background subtraction in pixel domain

* Details omitted in the talk

BlobNet Result

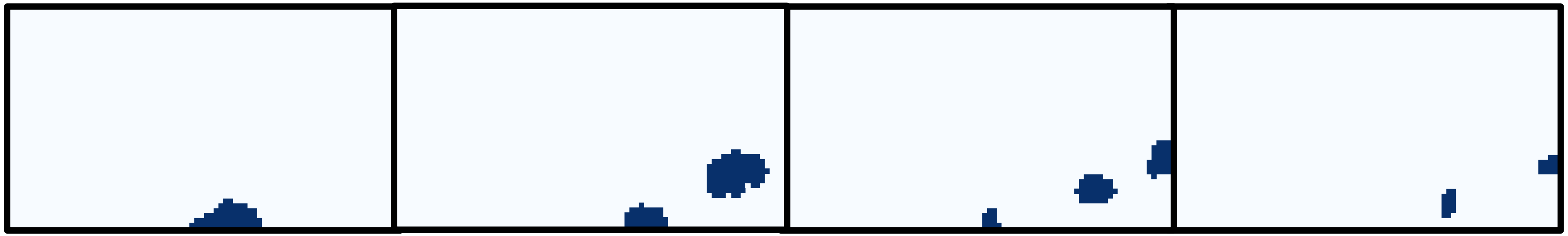
- *blob*: region where moving objects appear

Decoded Video

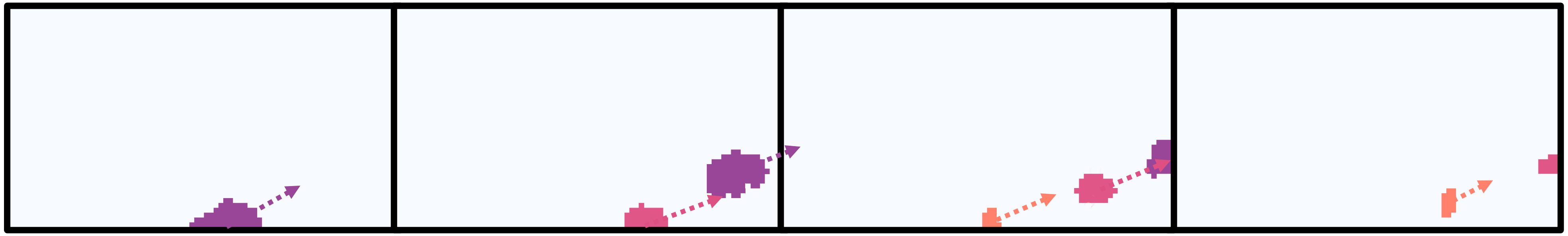
Detected *Blobs*

Detecting Tracks from Blobs

Blobs detected by BlobNet are not tracked yet



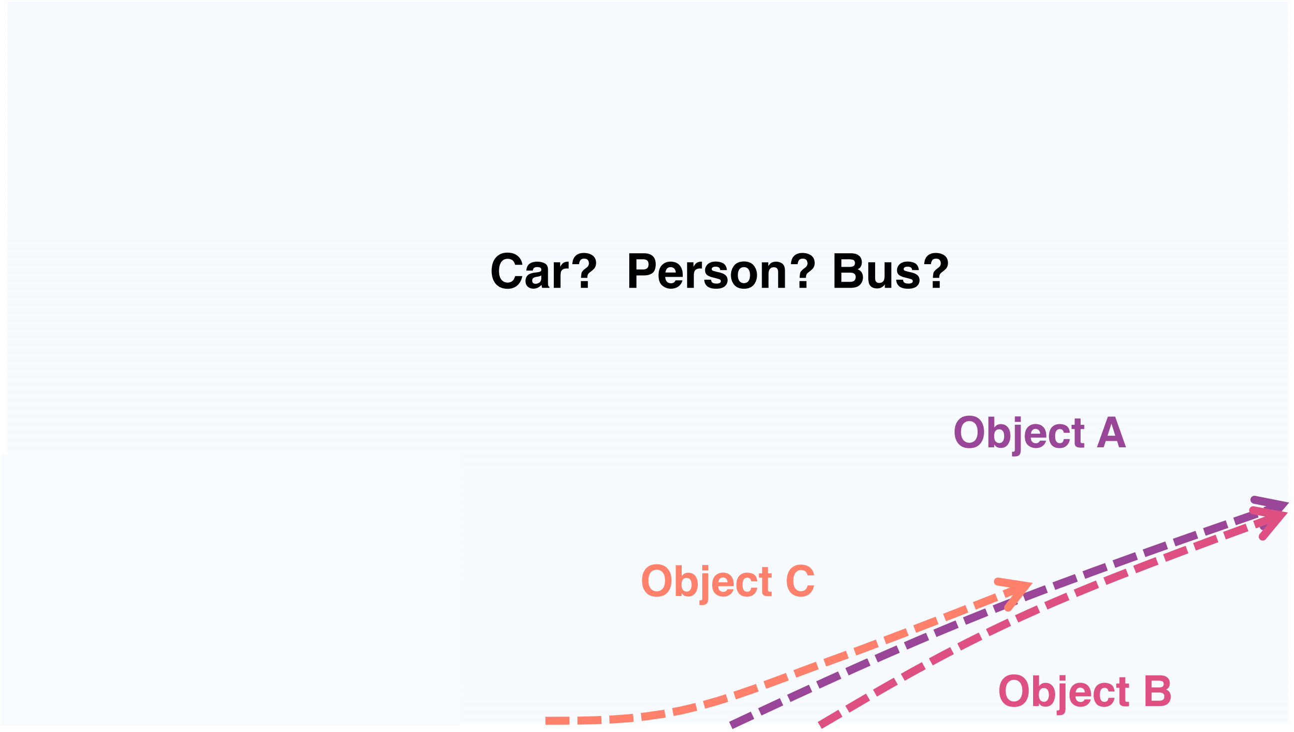
Tracking with **S**imple **O**nline and **R**ealtime **T**racking (SORT)



Frame Selection

Goal of Frame Selection

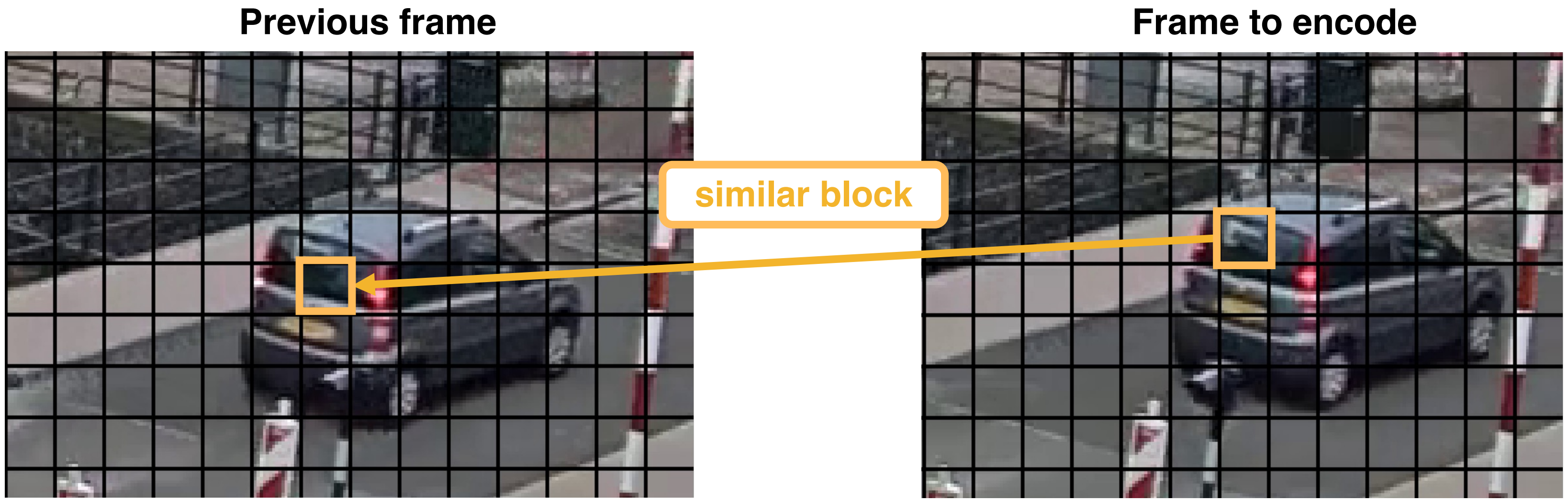
Goal: select minimal frames to decode

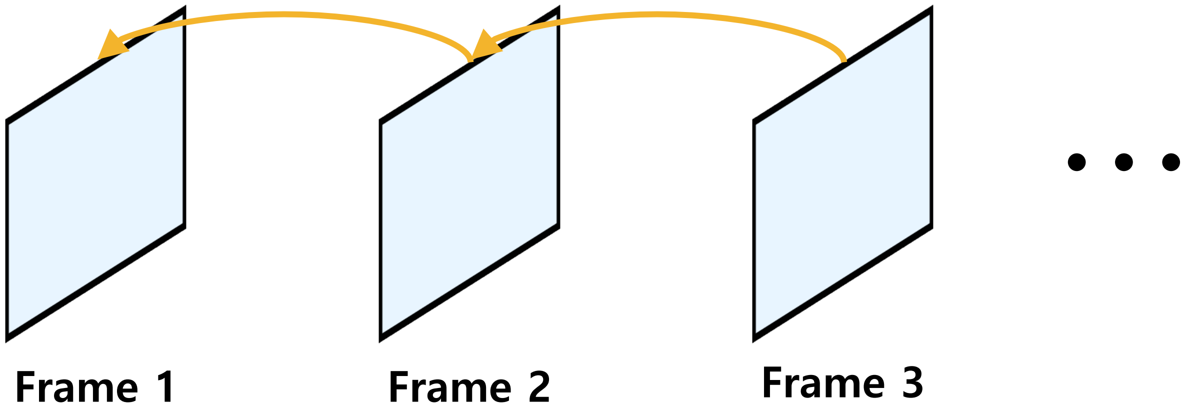


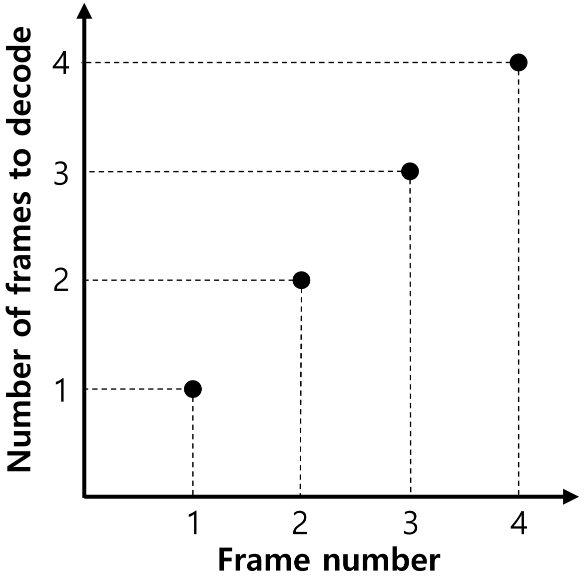
Decoding is required to see what ***kinds*** of objects they are

Can we just pick any of the frames to decode?

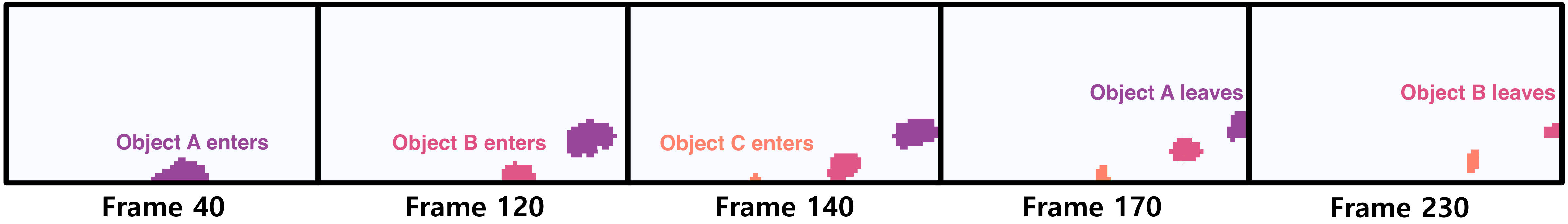
Not every frame has the same decoding cost

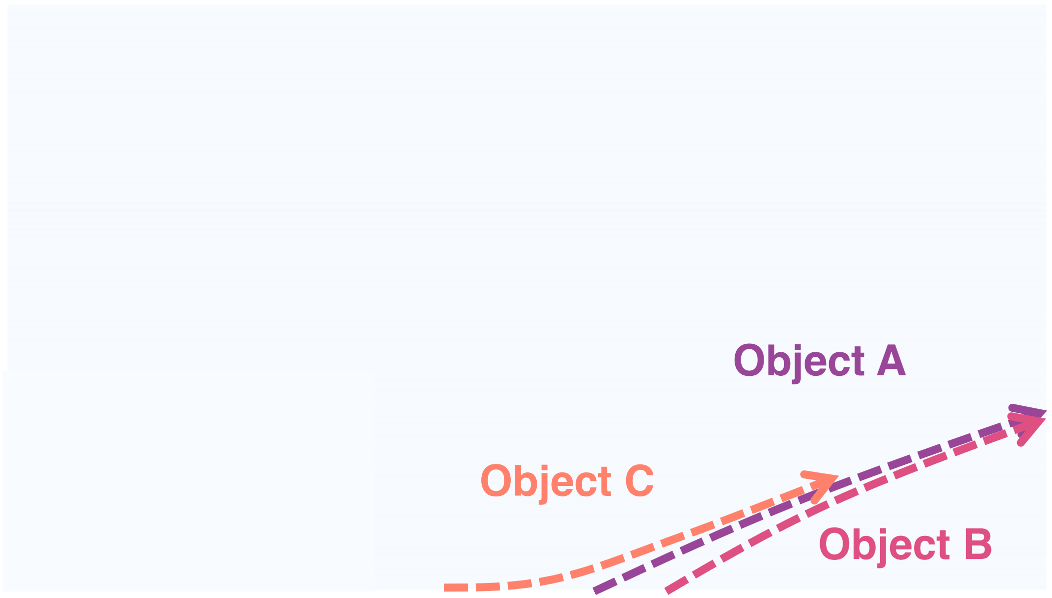


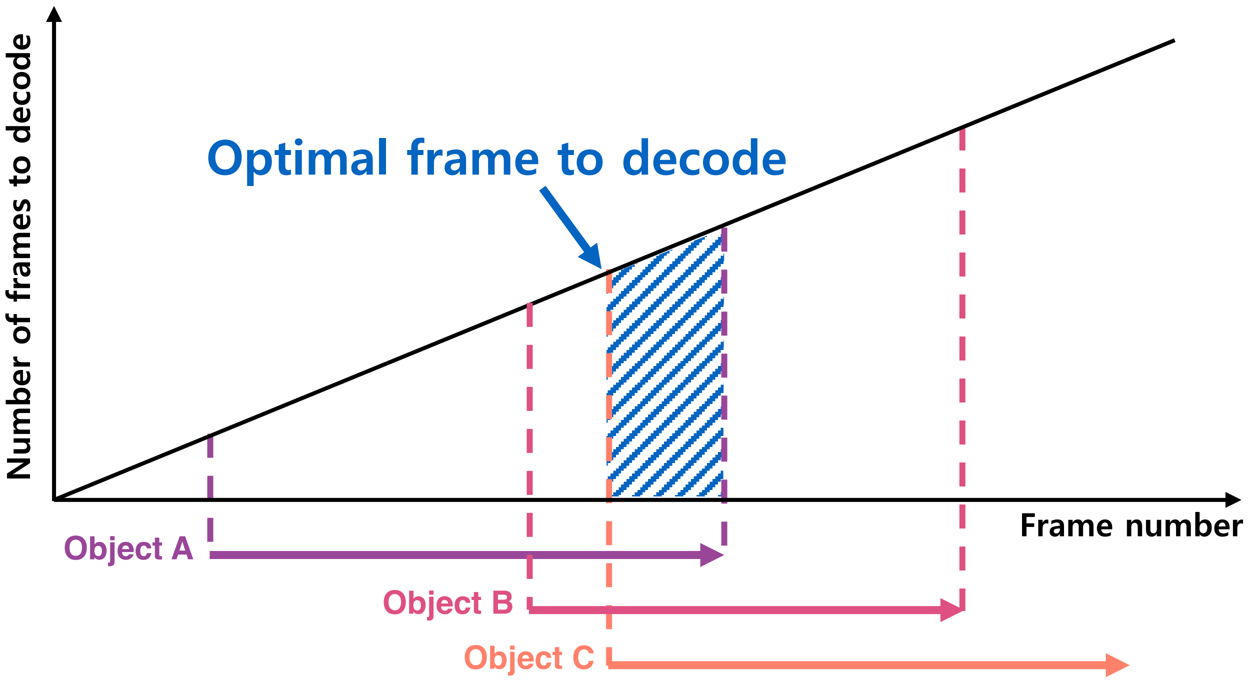




Dependency-Aware Frame Selection





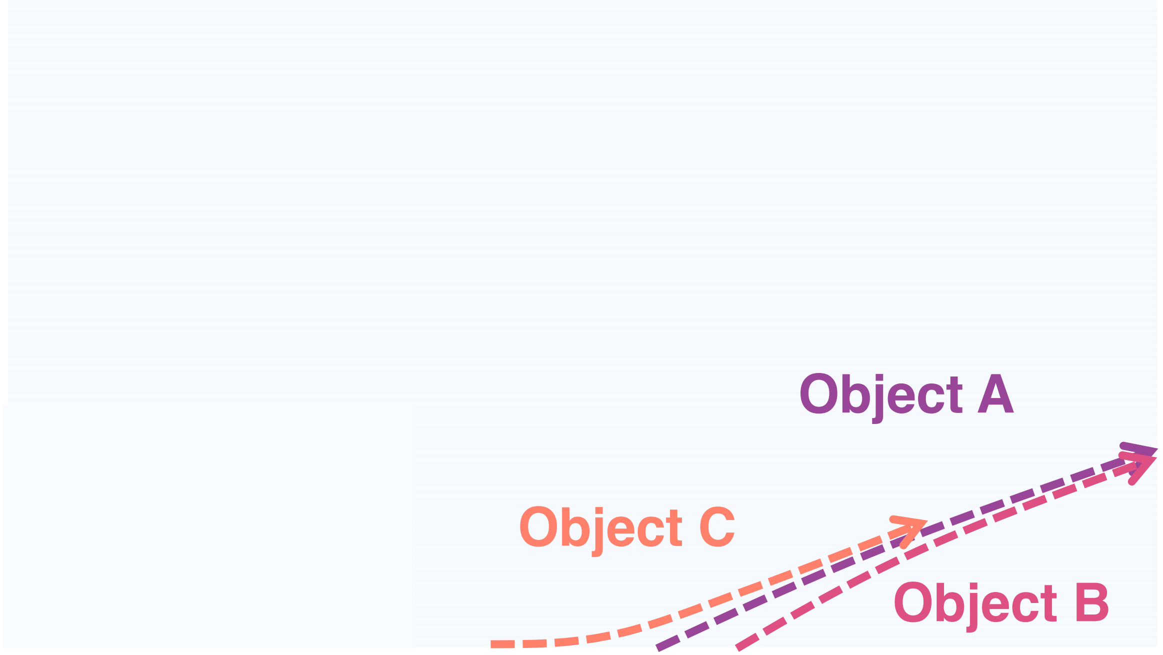


Decoding and Object Detection on Selected Frame

Label Propagation

Goal of Label Propagation

Track detection result

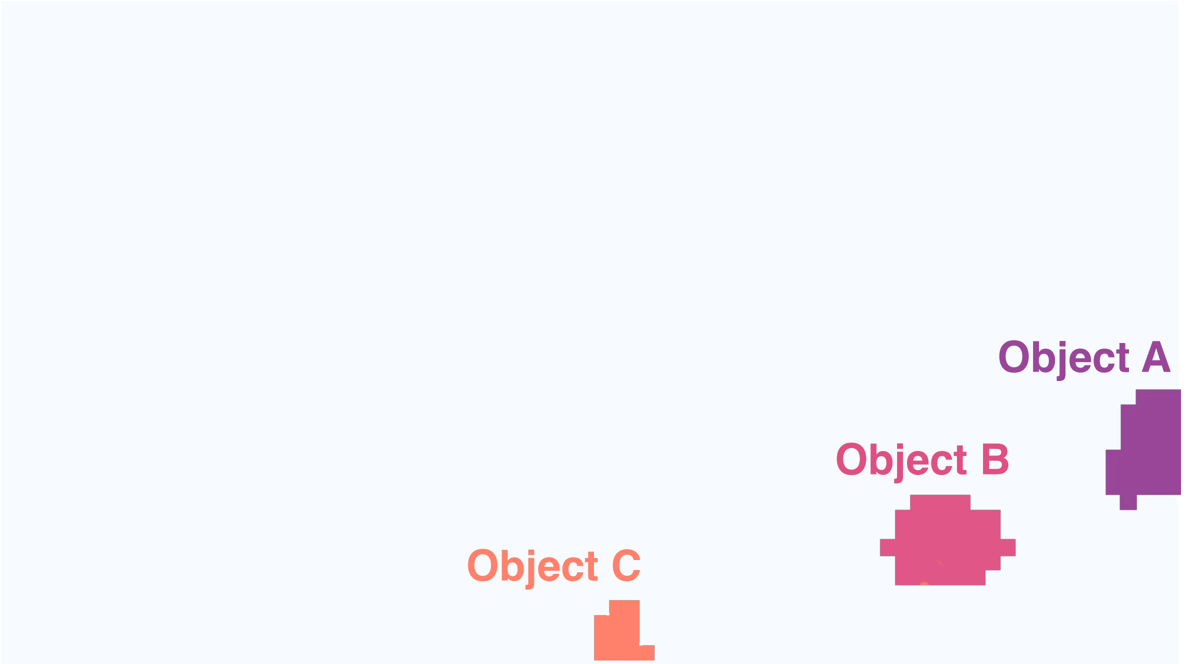


Object detection result

Goal: combine results from previous stages to label tracks

Overlap based label propagation

blobs at the same timestamp



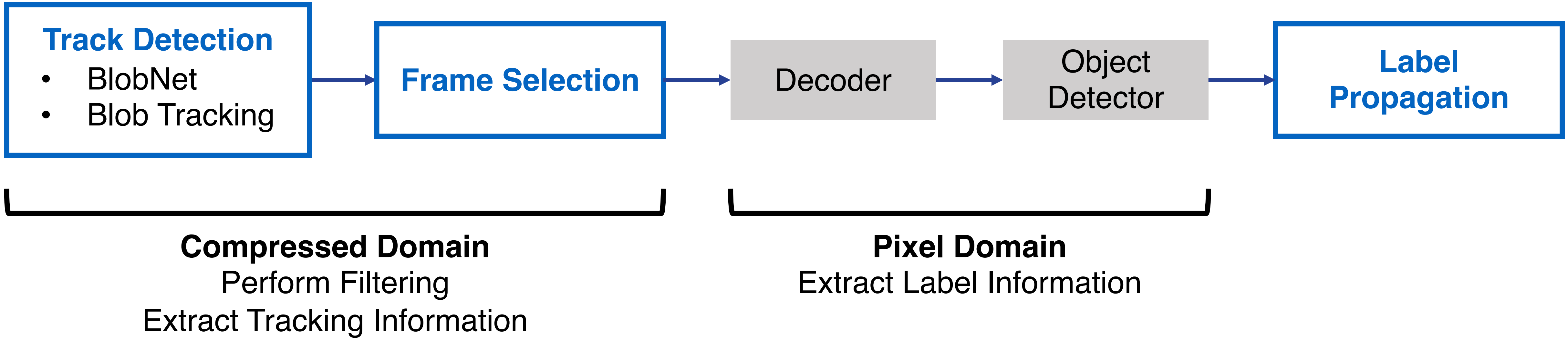
Object detection result

Retrieve *blob* location at the timestamp of object detected frame

Overlap based label propagation

Assigned labels are ***propagated*** throughout the track, including not decoded frames

CoVA Summary

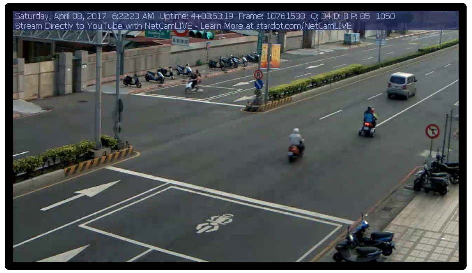


Evaluation Setup

Datasets: five live stream videos / Average 28 hours long



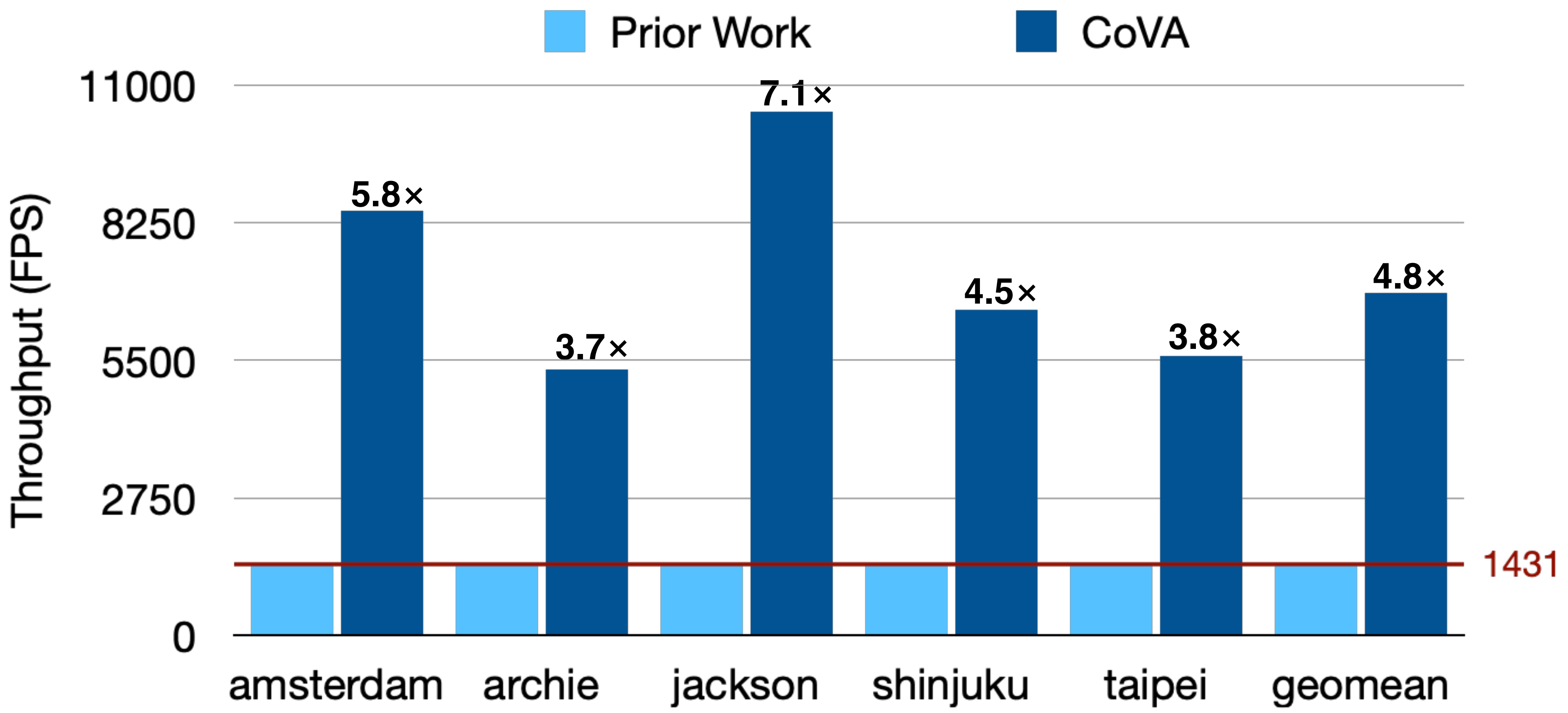




Query specification	
Binary Predicate (BP)	Frames where querying object appears
Global Count (CNT)	Average count of querying object
Local Binary Predicate (LBP)	BP with spatial constraint
Local Count (LCNT)	GC with spatial constraint

System specification	
Software	C++ & Rust / CUDA 11.5
Decoder	FFmpeg v4.41 / NVDEC v5
CPU	Two Intel Xeon CPU Gold 6226R
GPU	NVIDIA RTX 3090

End to End System Throughput Improvement



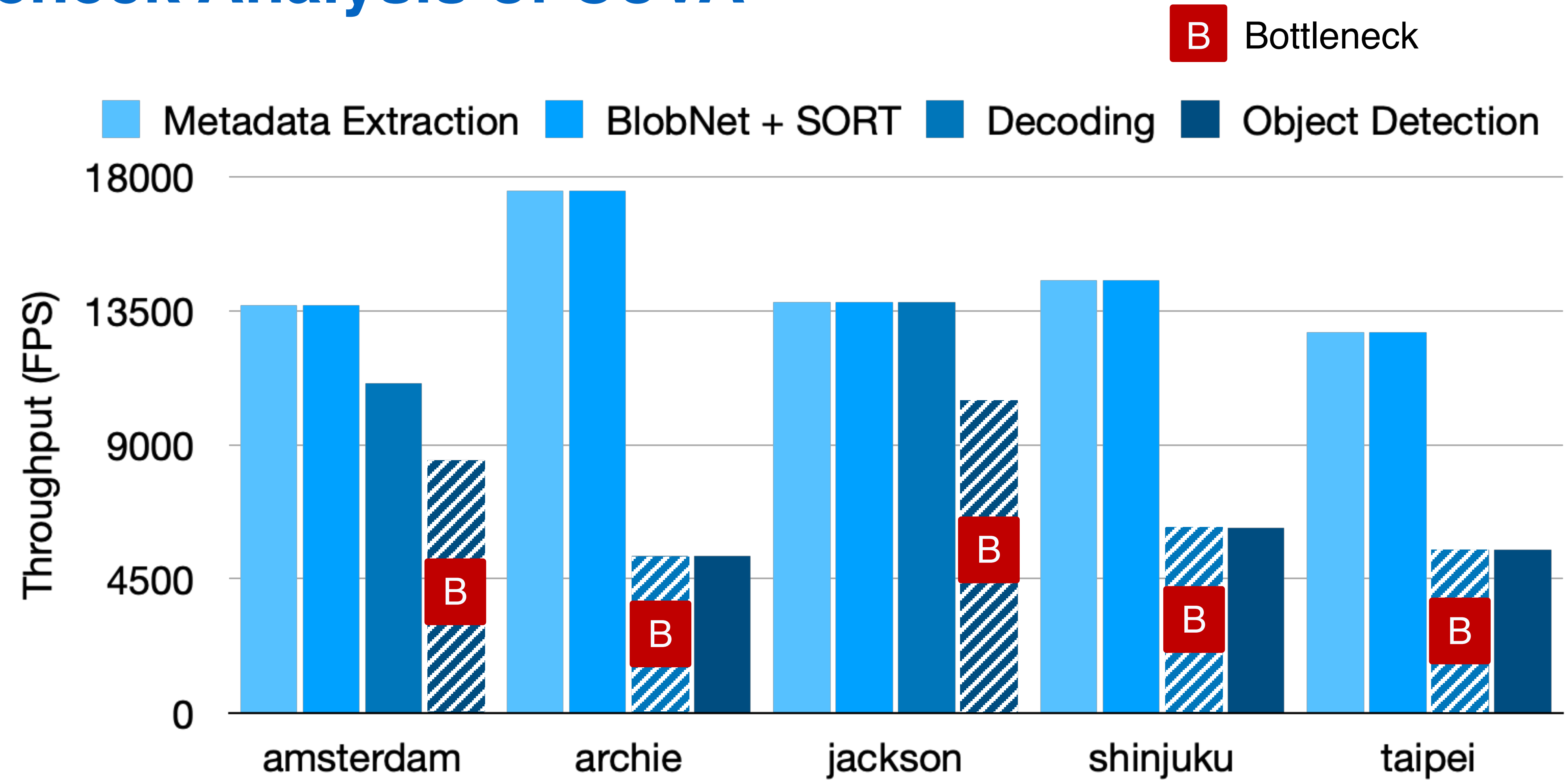
Achieves 4.8× higher throughput in average compared to prior work

Filtration rate

Dataset	Decode Filtration Rate (%)	Inference Filtration Rate (%)
amsterdam	87.16	99.60
archie	72.94	99.15
jackson	94.81	99.79
shinjuku	77.18	99.26
taipei	74.03	99.81
geomean	80.80	99.39

Reduces decoding workload by 80.8%, and inference by 99.4% on average

Bottleneck Analysis of CoVA



Bottleneck of varies across dataset

Compressed domain filtering never becomes the bottleneck

Implication on accuracy

Dataset	BP (%)	CNT (Err)	Ground Truth*
amsterdam	85.79	0.15	1.40
archie	86.96	0.04	0.16
jackson	86.13	0.10	0.56
shinjuku	90.15	0.30	2.18
taipei	87.74	1.10	5.03
geomean	87.34		

*Comparison made against YOLOv4 as ground truth

Degrades accuracy in modest level comparable to prior works

E.g., Degradation in binary predicate query is in the range of 10-15%

Conclusion

- Novel video analytics pipeline that introduces compressed domain analysis
- 4.8× on average speedup by addressing decoding bottleneck
- Support for spatial query

Opensourced
Artifact evaluated

QnA