PacketGame:

Multi-Stream Packet Gating for Concurrent Video Inference at Scale

Mu Yuan, Lan Zhang, Xuanke You, Xiang-Yang Li

University of Science and Technology of China

Outline

- **Background**
- PacketGame Design
- Evaluation

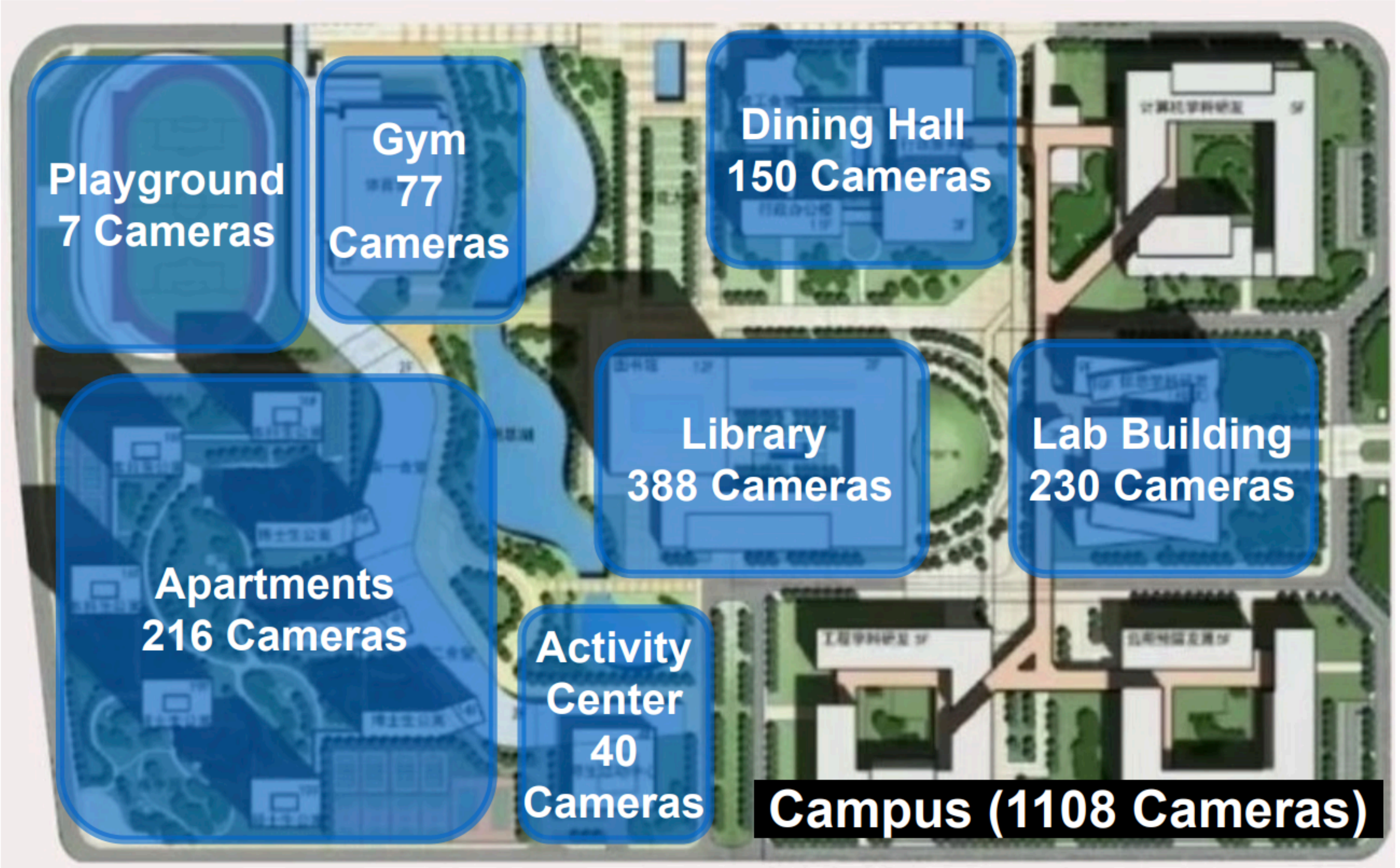
Background
Development Experience

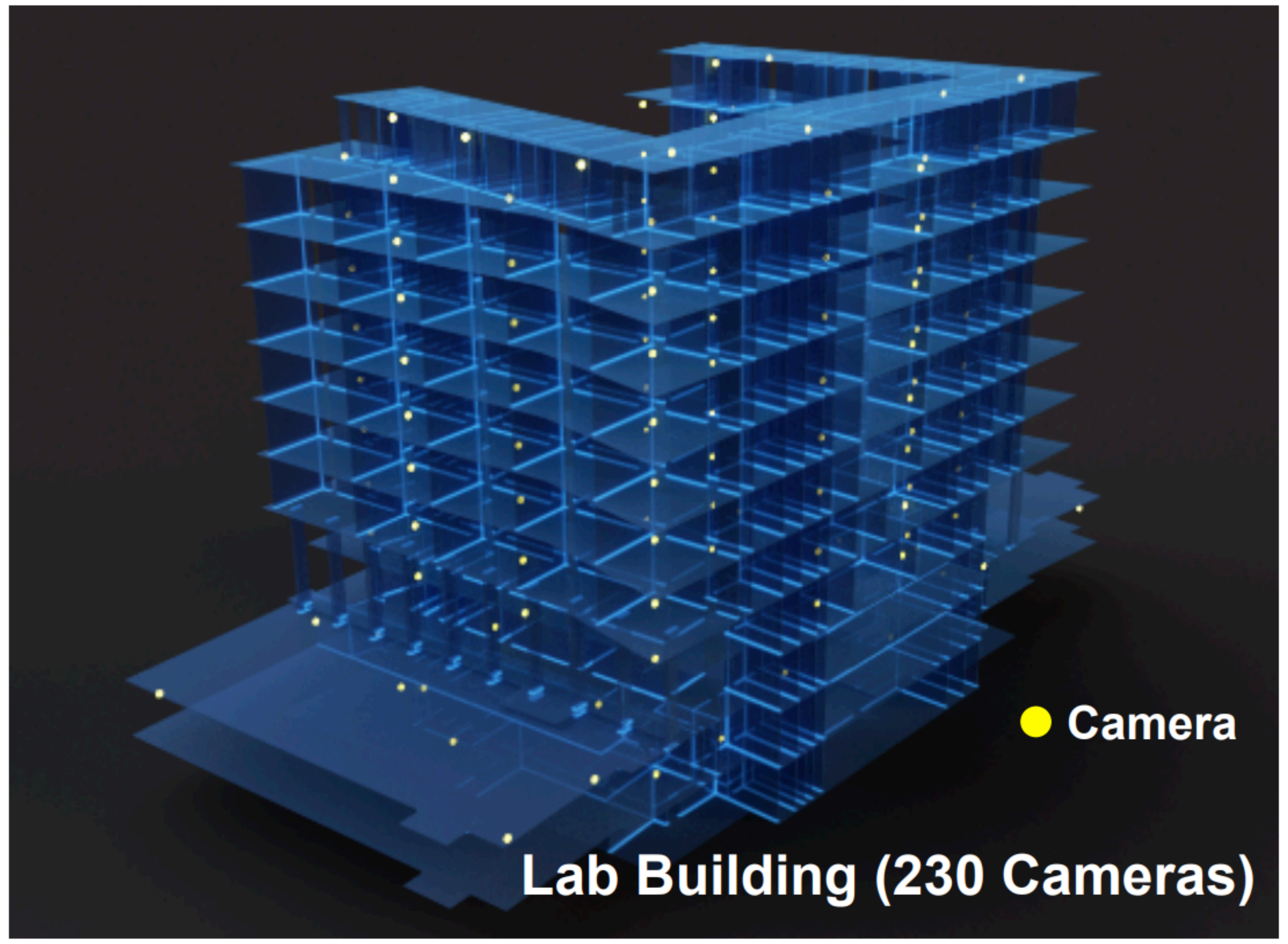
- Video analytics system at University of Science and Technology of China
  - mobility analysis and anomaly detection

Background

Development Experience

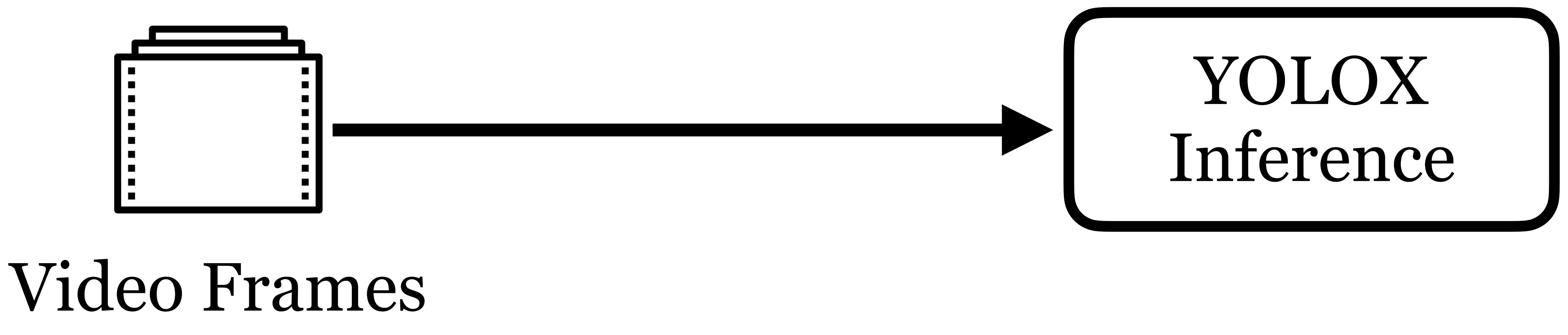
- Video analytics system at University of Science and Technology of China
  - mobility analysis and anomaly detection
  - 1108 real-time 1080p streams from IP cameras

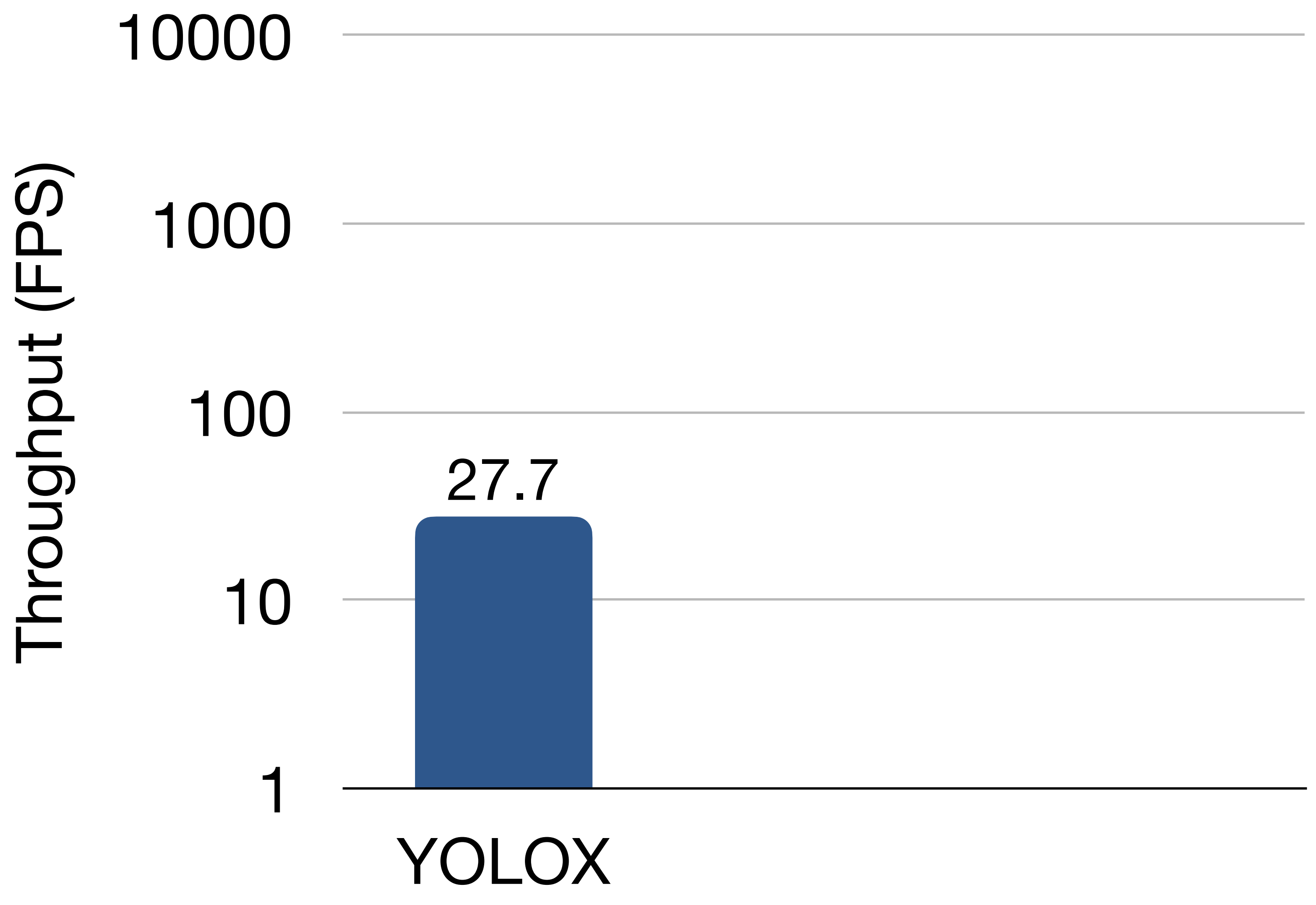




Background
Development Experience

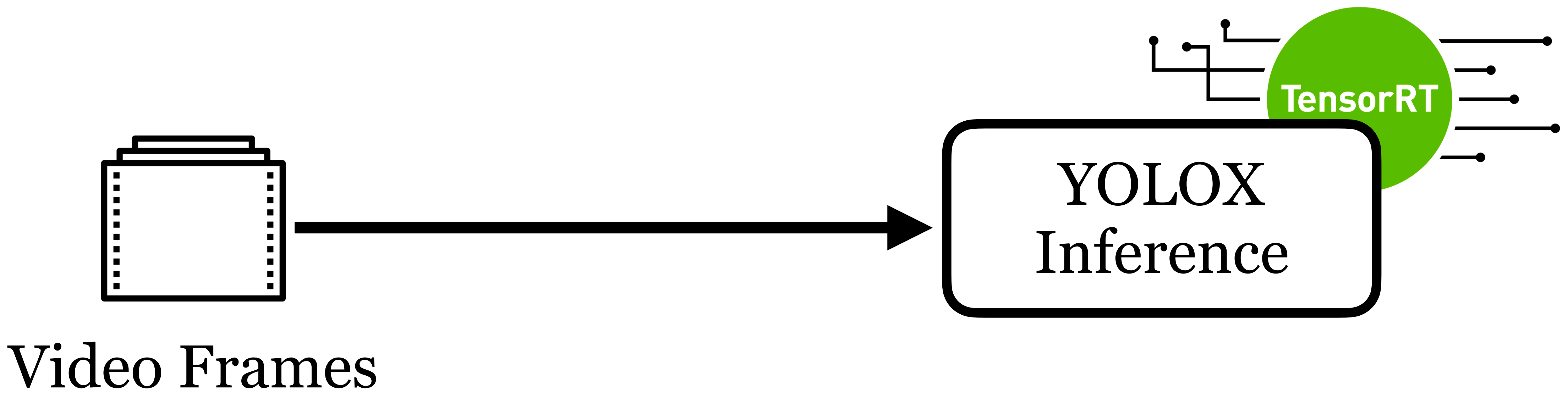
- System Setup
  - 12 CPUs + TITAN X GPU edge server
  - YOLOX for object detection (on GPU)
    Ge, Zheng, et al. "Yolox: Exceeding yolo series in 2021." *arXiv preprint arXiv:2107.08430* (2021).

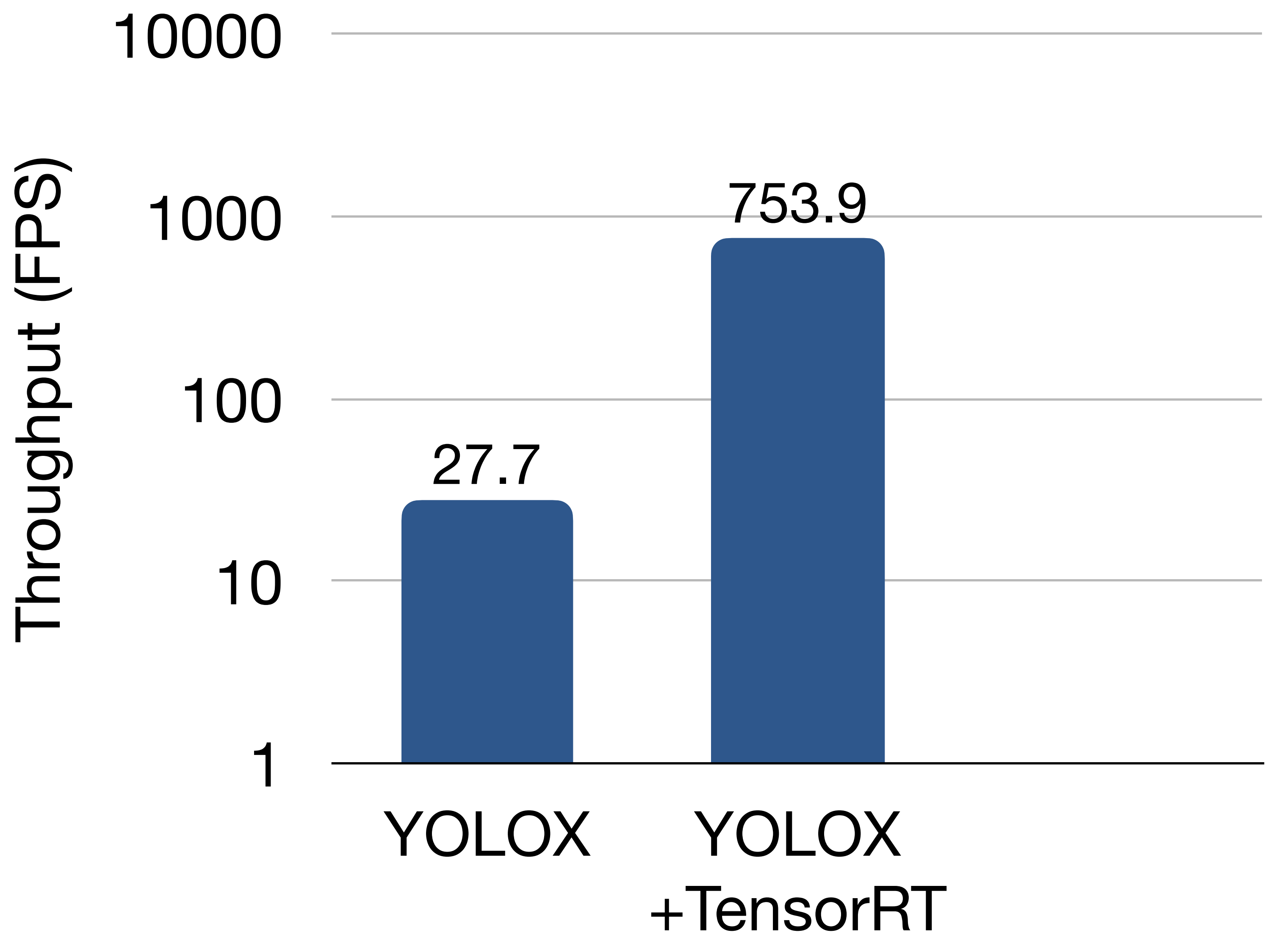




Background
Development Experience

- Applying optimization techniques
  - NVIDIA TensorRT (model inference acceleration)



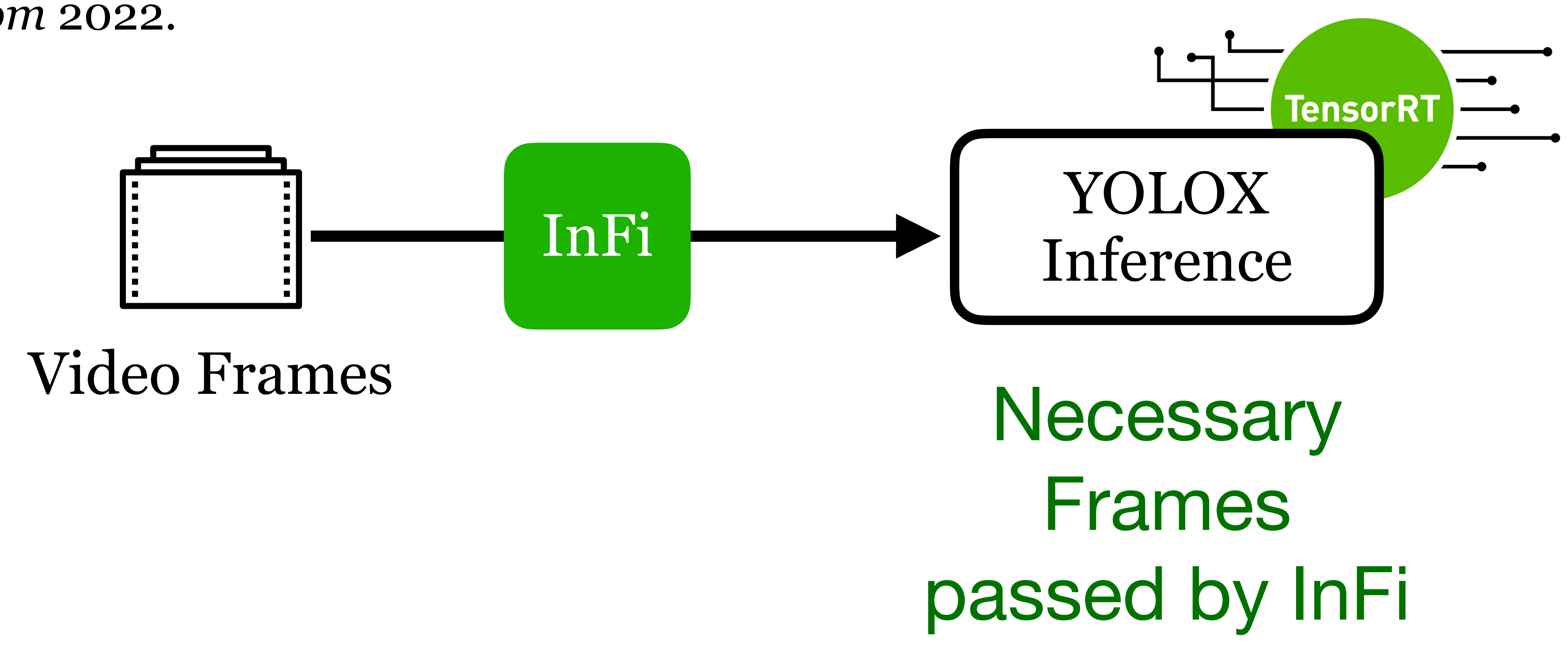


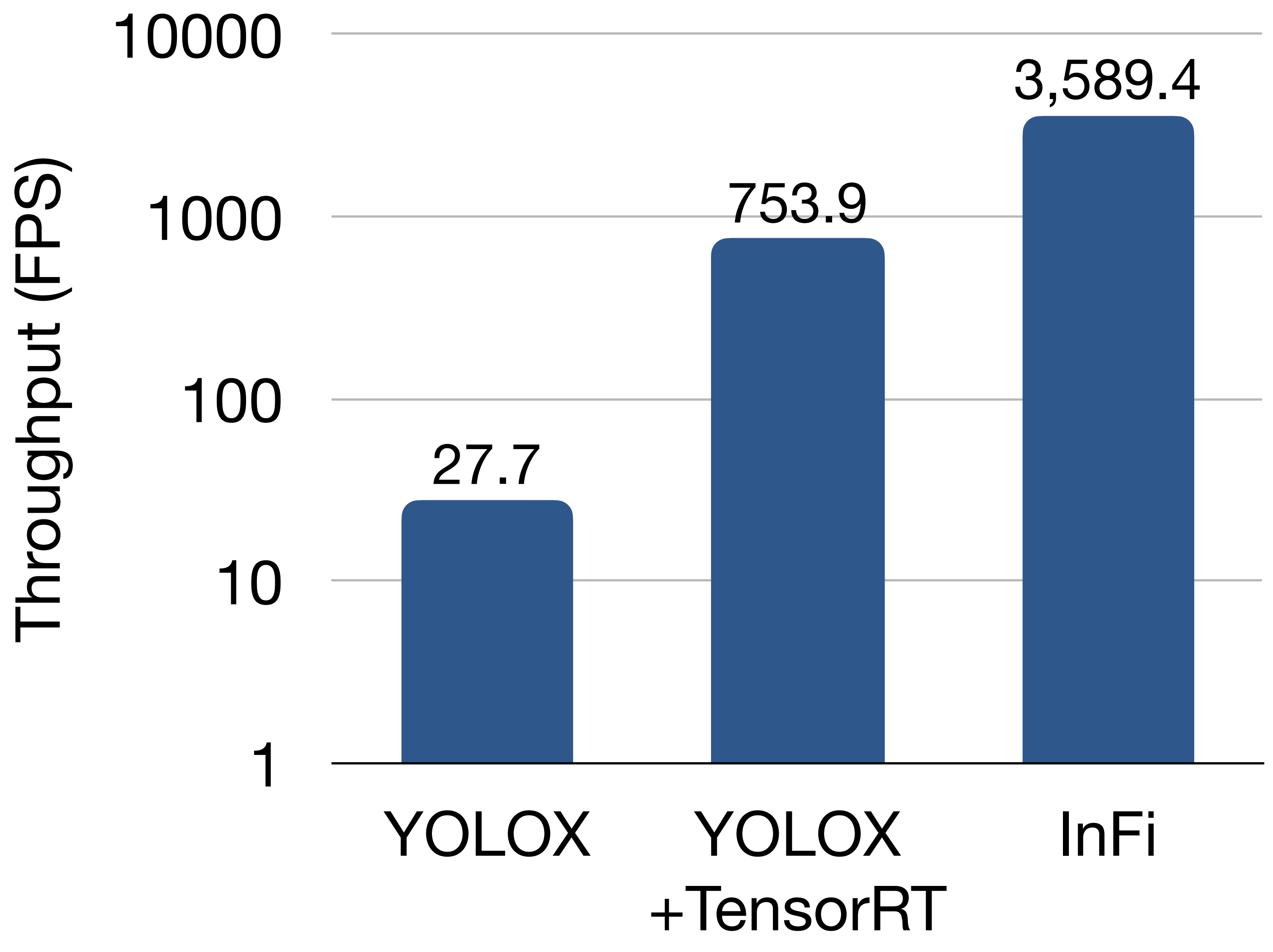
Background

Development Experience

- Applying optimization techniques
  - NVIDIA TensorRT (model inference acceleration)
  - InFi (frame filtering, our MobiCom’22 paper)

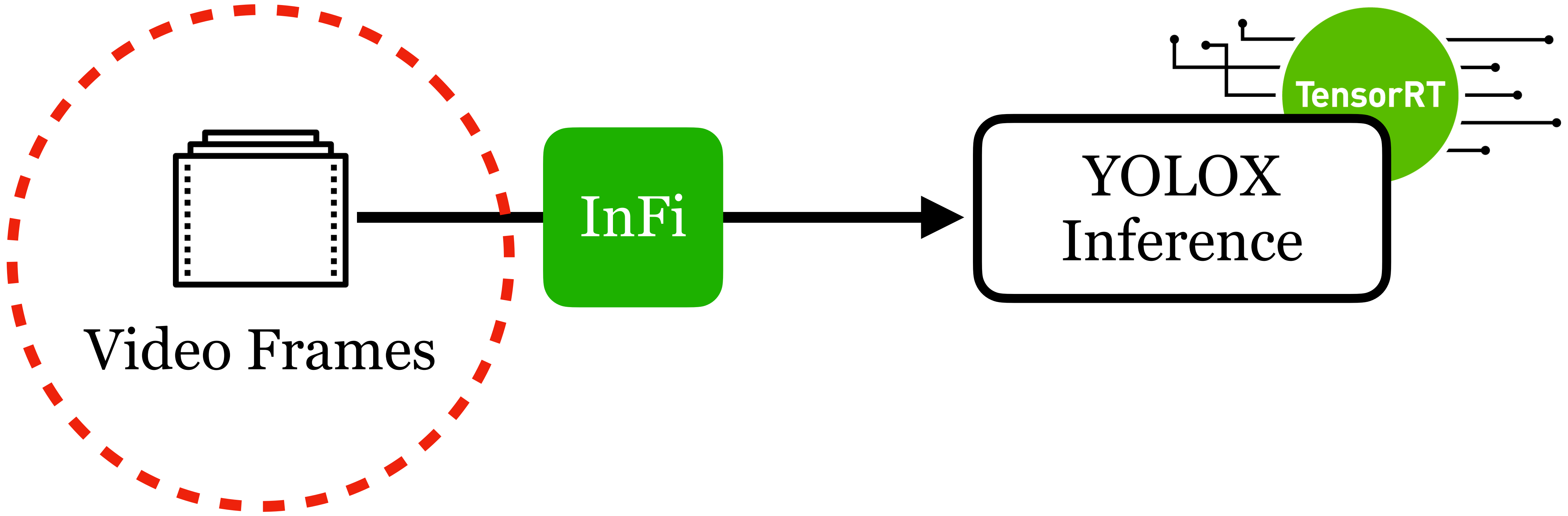
    Yuan, Mu, et al. "Infi: end-to-end learnable input filter for resource-efficient mobile-centric inference." *MobiCom* 2022.





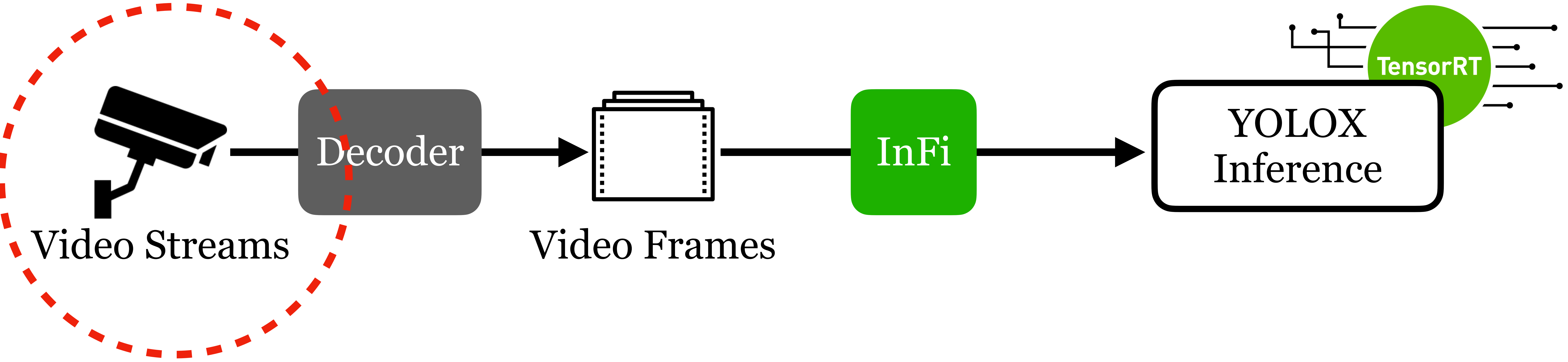
Background
Concurrency Bottleneck

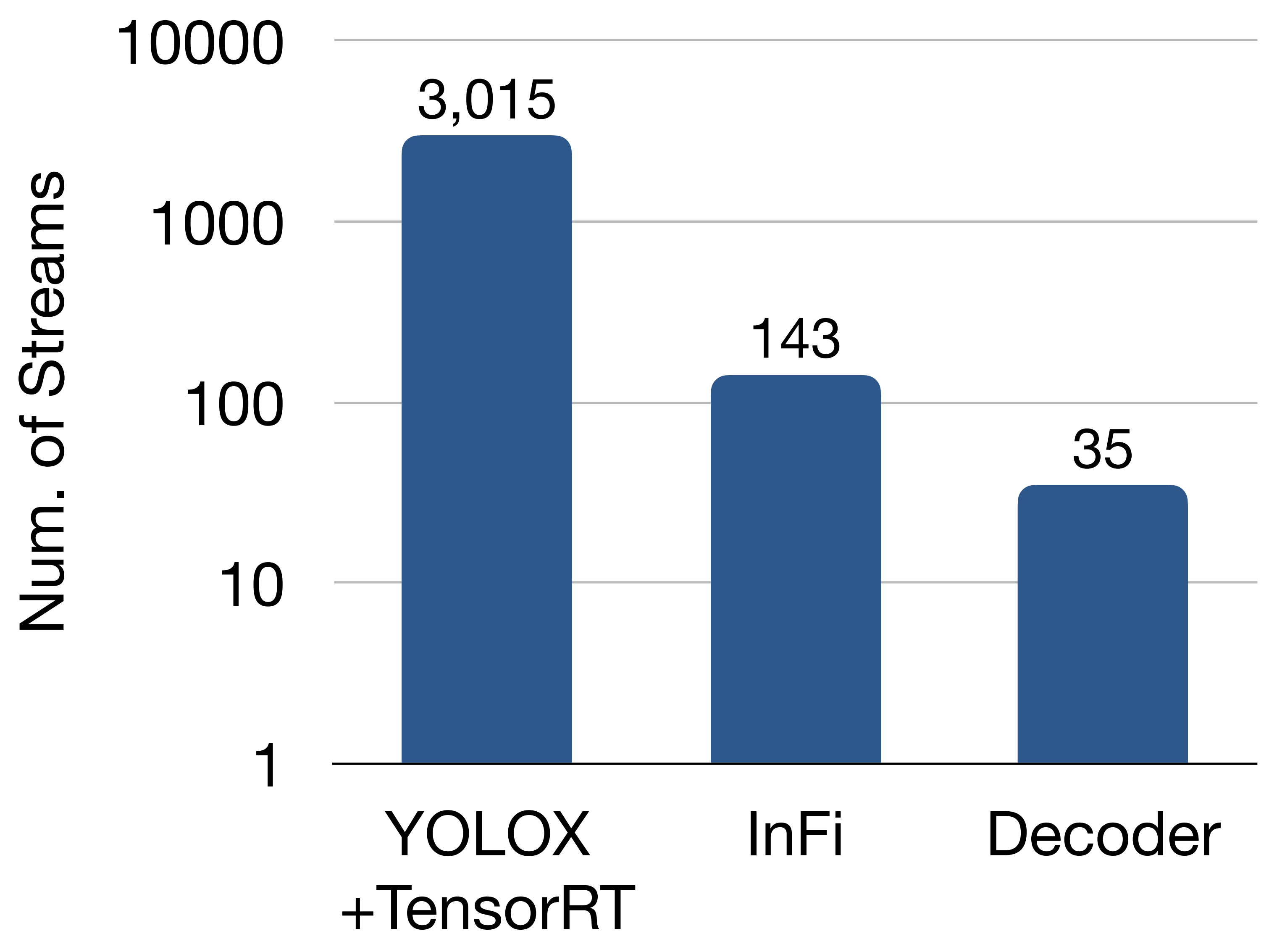
- Concurrency benchmarks



Background
Concurrency Bottleneck

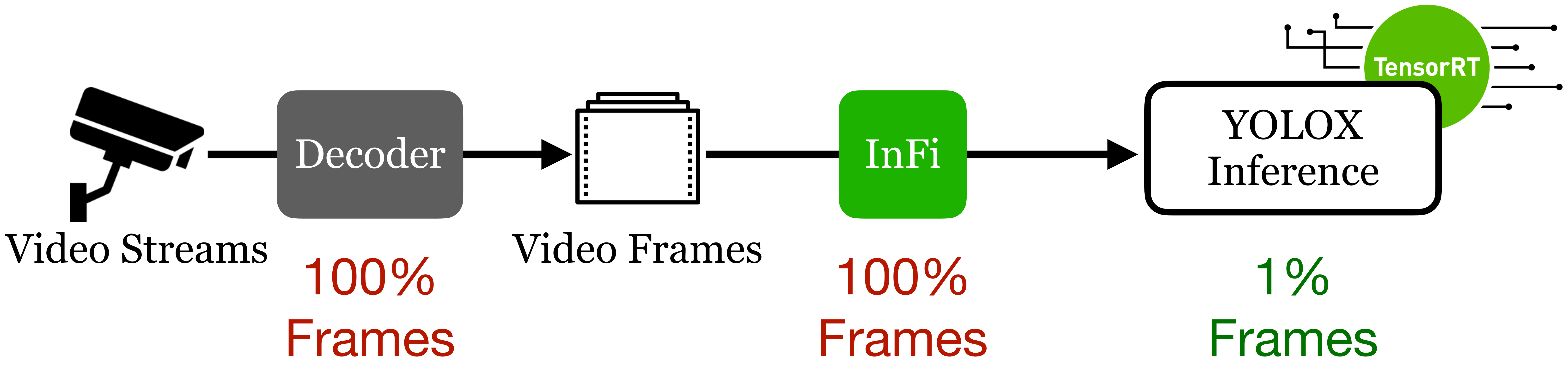
- Concurrency benchmarks
  - End-to-end concurrency is bottlenecked by the decoder (on 12 CPUs)

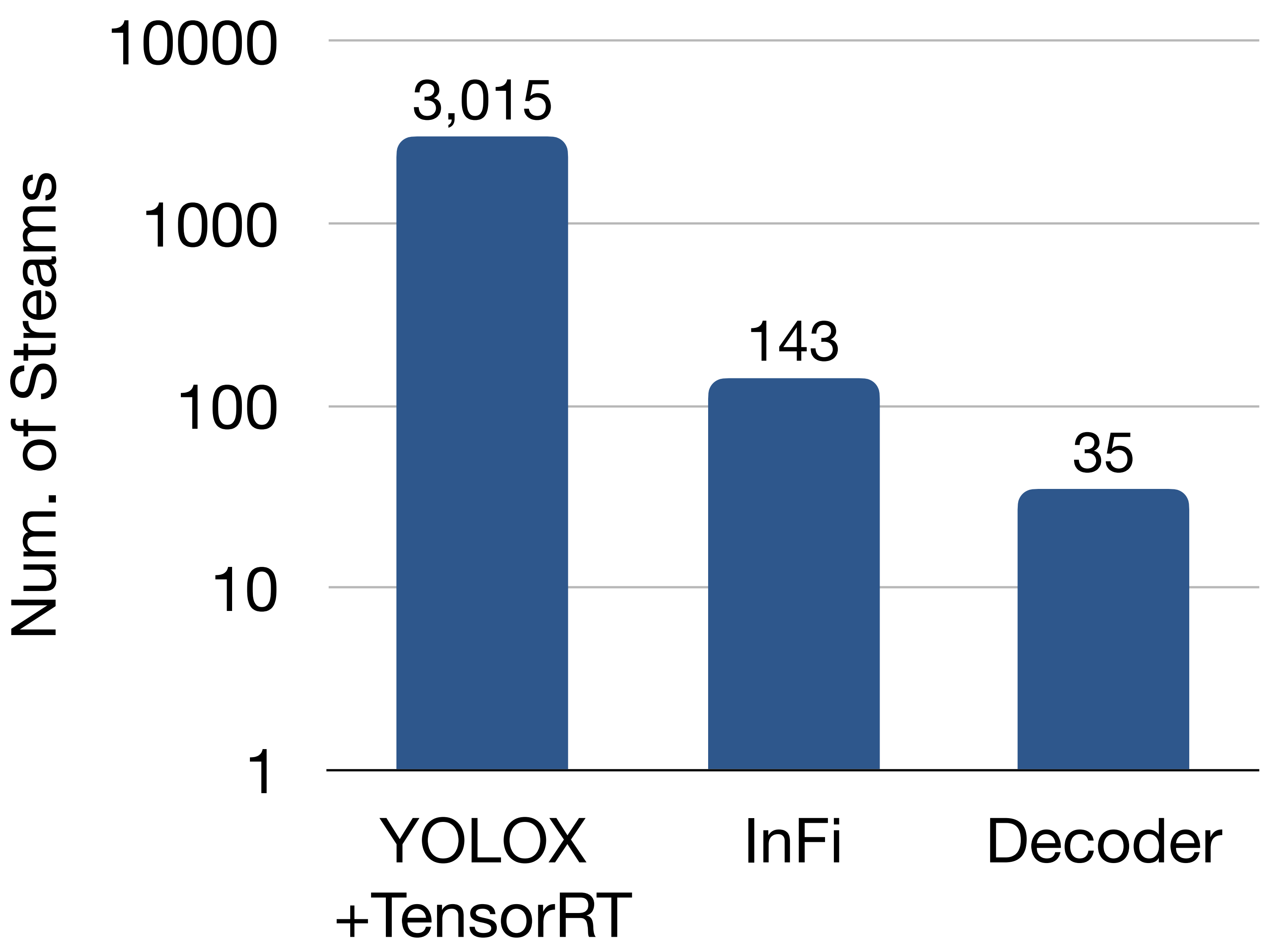




Background
Concurrency Bottleneck

- Concurrency benchmarks
  - End-to-end concurrency is bottlenecked by the decoder (on 12 CPUs)
  - Reason: all-frame decoding vs. partial inference

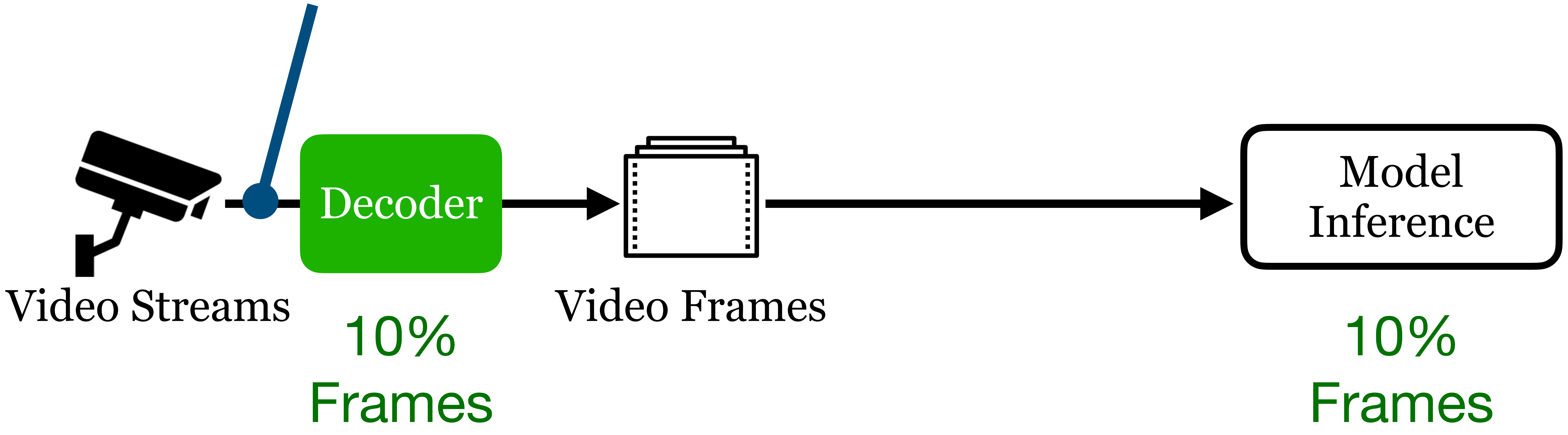




Background

New Idea

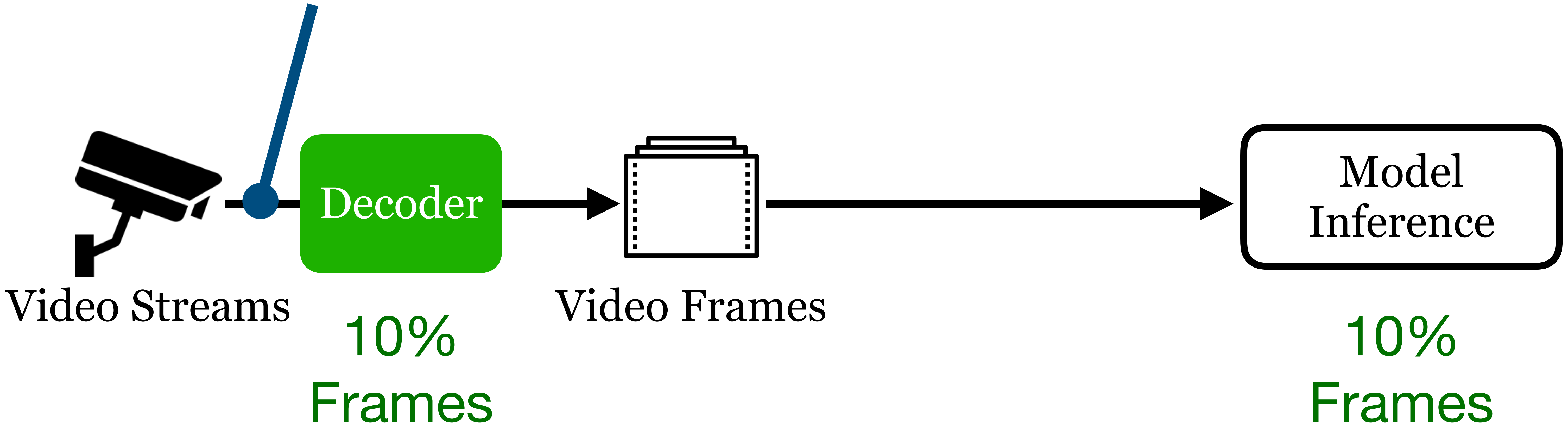
- Packet gating
  - selectively passing video packets to the decoder
  - reducing both decoder and inference overheads



Background

New Idea

- Packet gating
  - selectively passing video packets to the decoder
  - reducing both decoder and inference overheads



Comparison with Existing Ideas

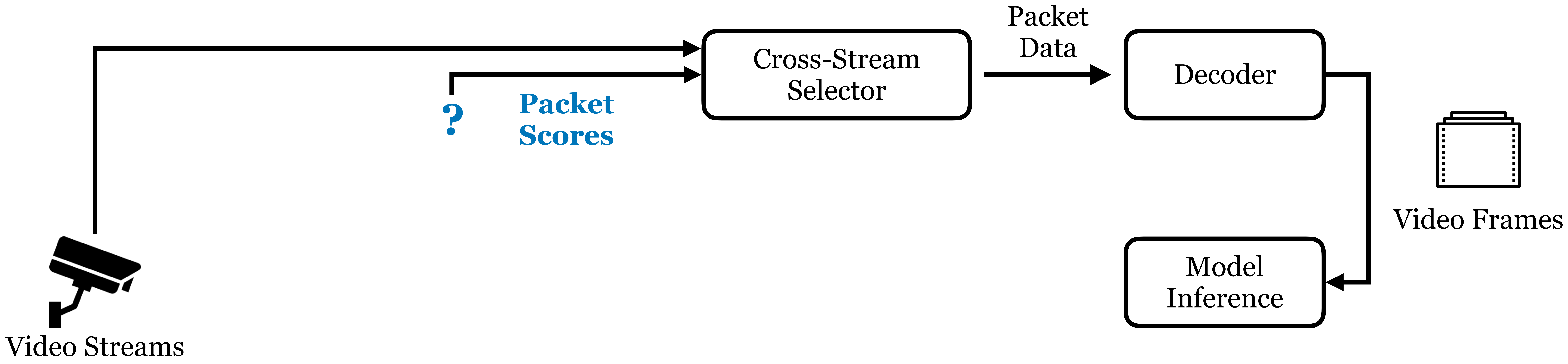
Methods	Reduce Decode	Commodity Cameras	Offline Videos
Video Compression	✓	✗	✗
On-Camera FF	✓	✗	✗
On-Server FF	✗	✓	✓
Model Acceleration	✗	✓	✓
Packet Gating	✓	✓	✓

Outline

- Background
- **PacketGame Design**
- Evaluation

PacketGame Design

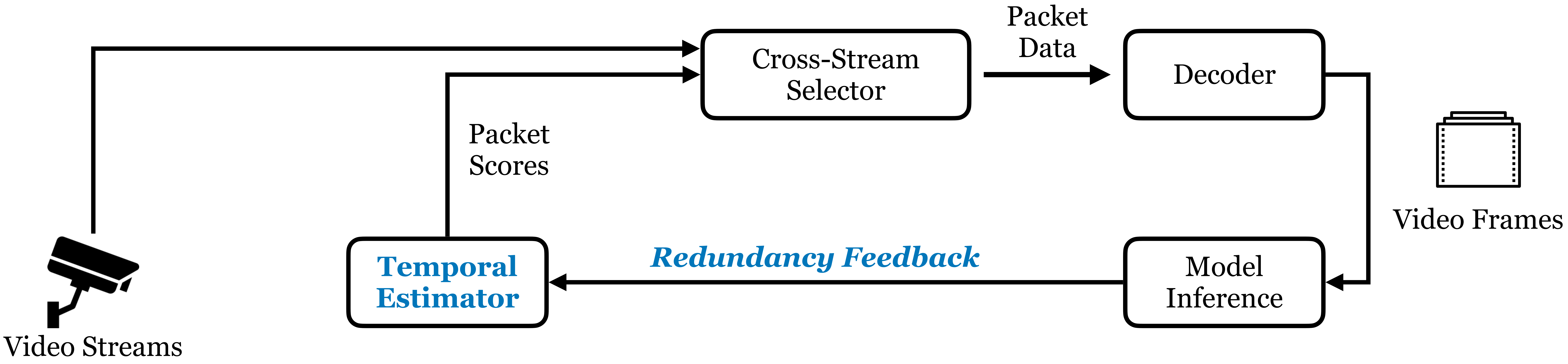
- To selectively pass packets to the decoder, we need **quantitive "scores"** for video packets from concurrent streams



PacketGame Design

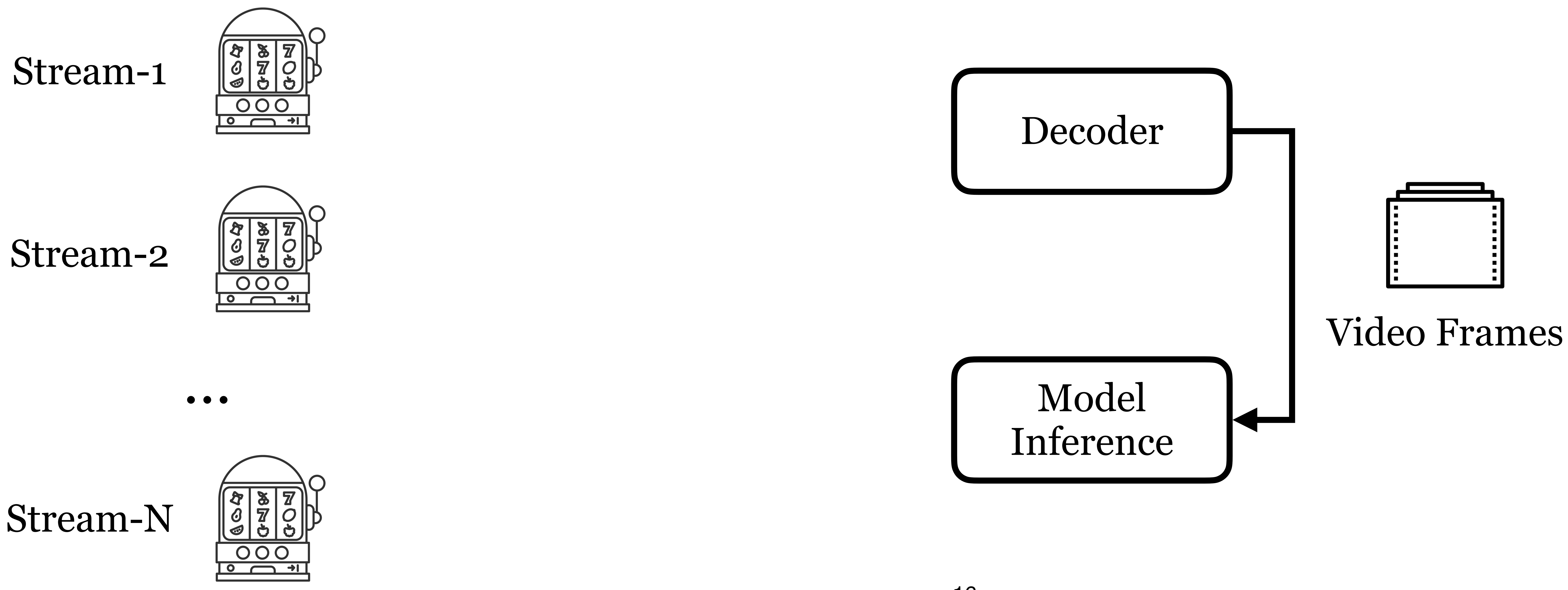
Temporal Estimator

- Available hint#1: historical feedback
    - Redundancy: the new inference result == the latest result



PacketGame Design
Temporal Estimator

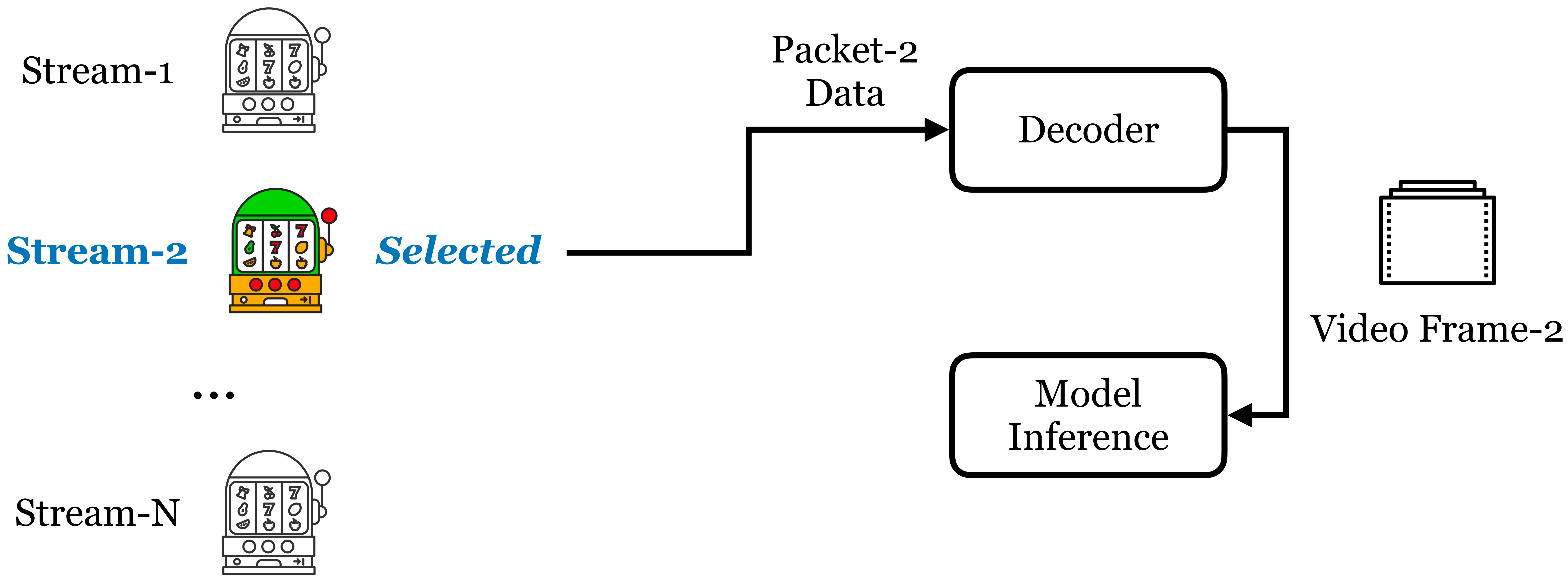
- Available hint#1: historical feedback
  - MAB-based approach



PacketGame Design

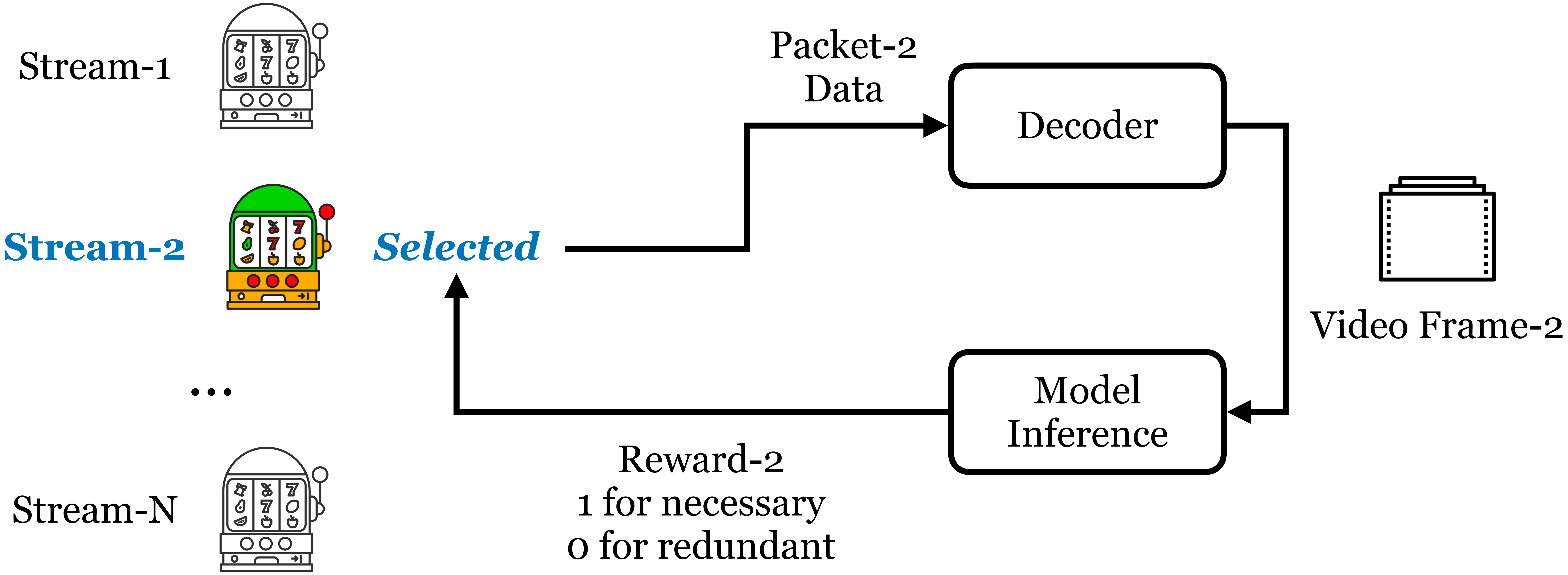
Temporal Estimator

- Available hint#1: historical feedback
  - MAB-based approach



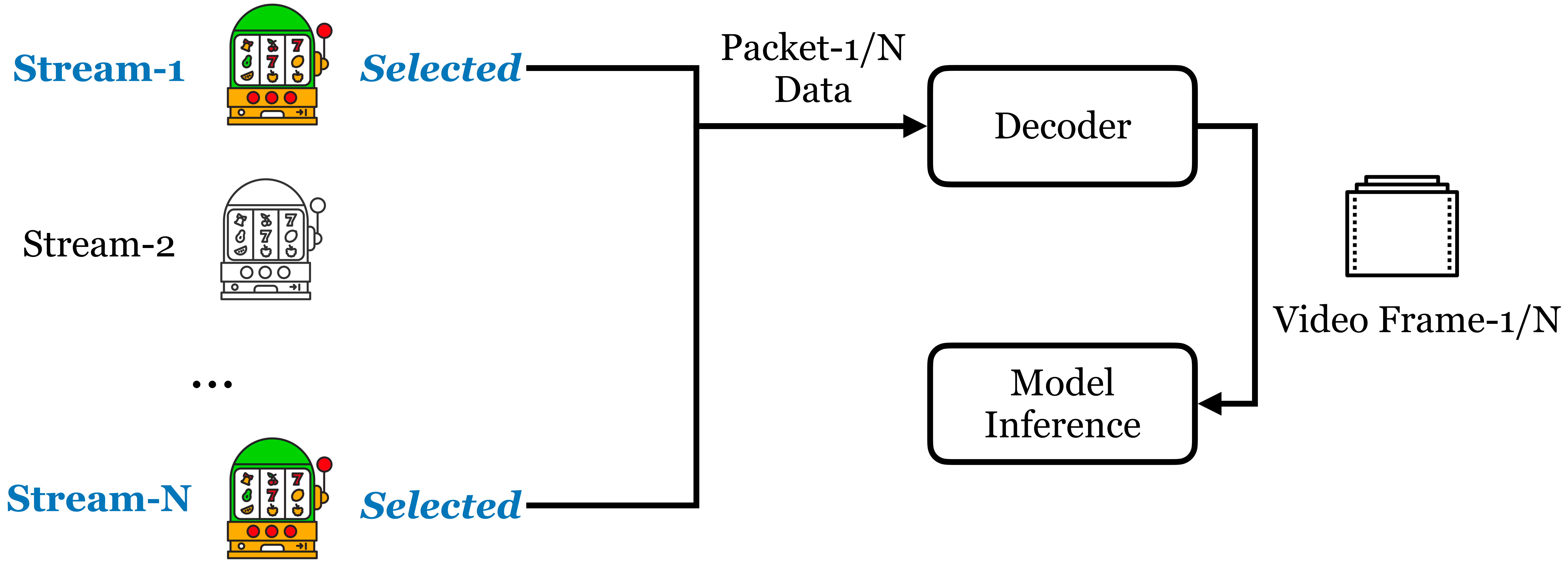
PacketGame Design
Temporal Estimator

- Available hint#1: historical feedback
  - MAB-based approach



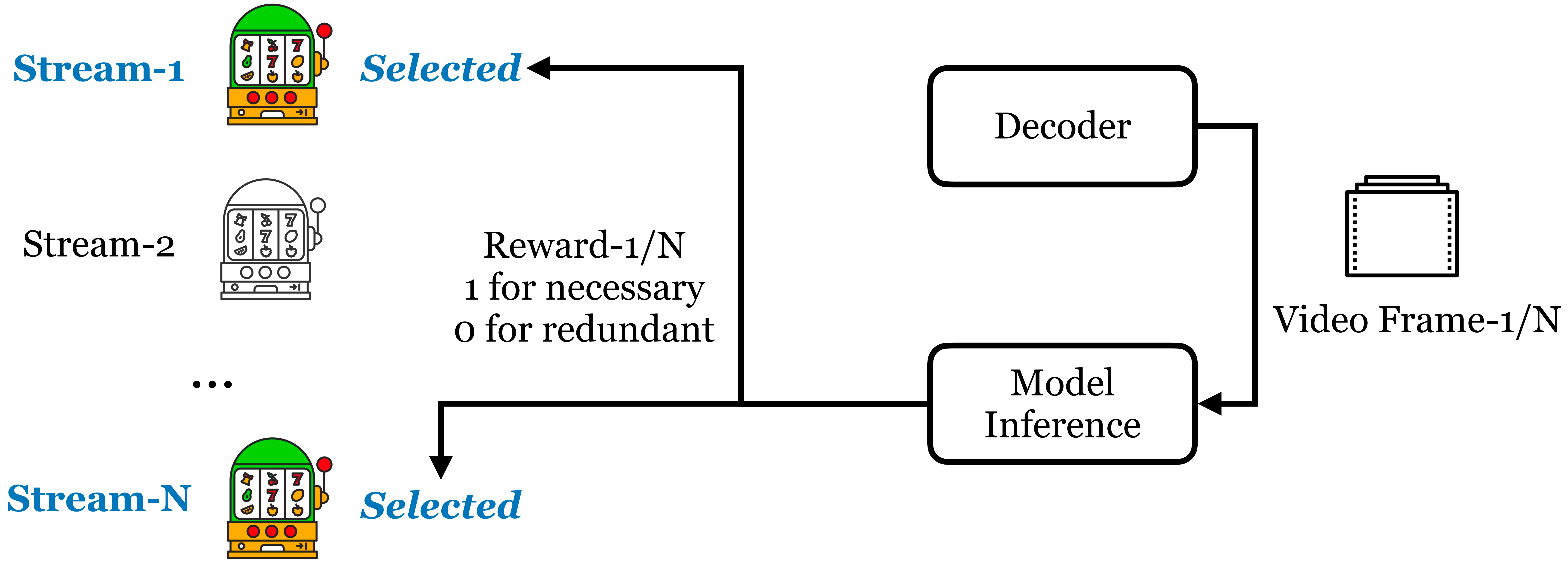
PacketGame Design
Temporal Estimator

- Available hint#1: historical feedback
  - MAB-based approach



PacketGame Design
Temporal Estimator

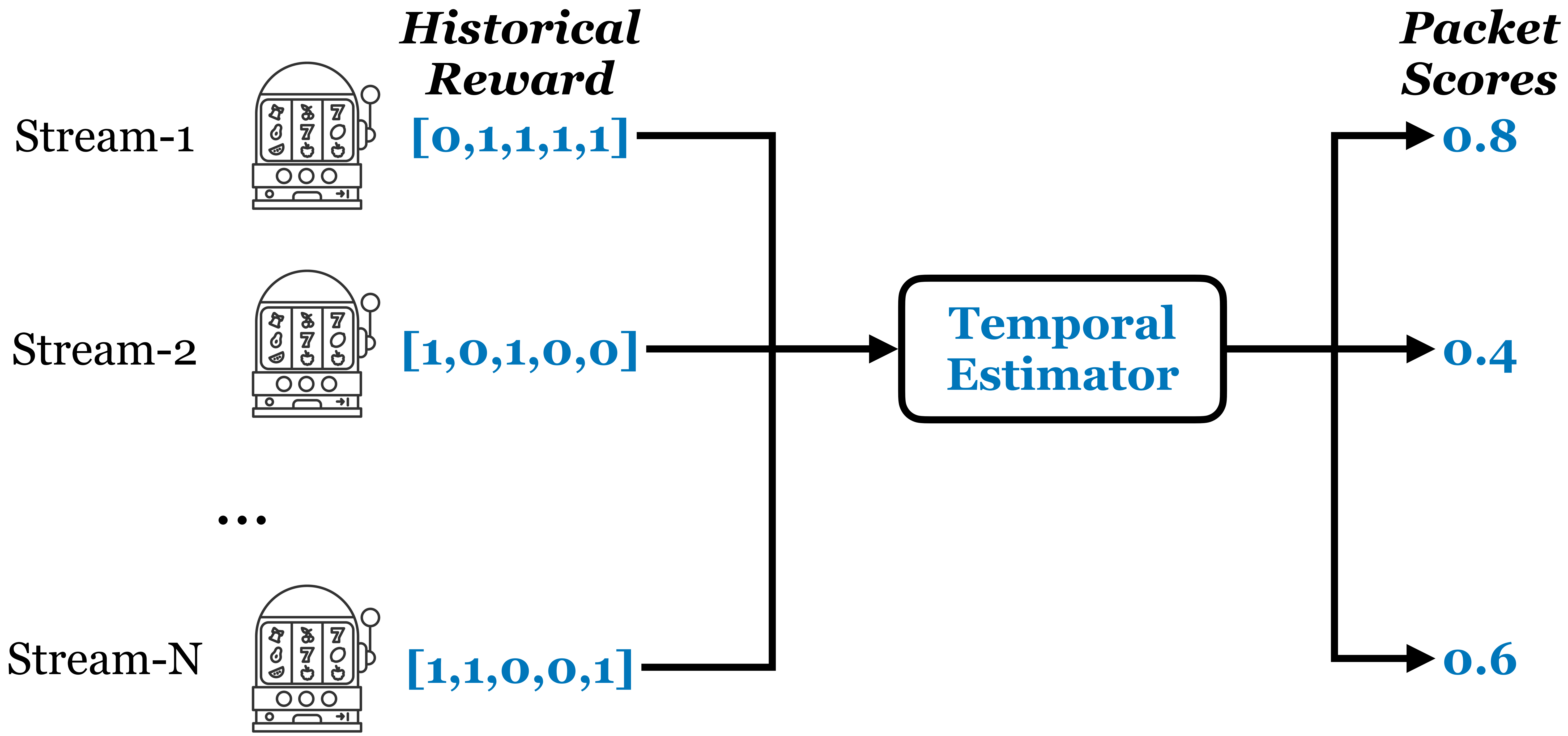
- Available hint#1: historical feedback
  - MAB-based approach



PacketGame Design

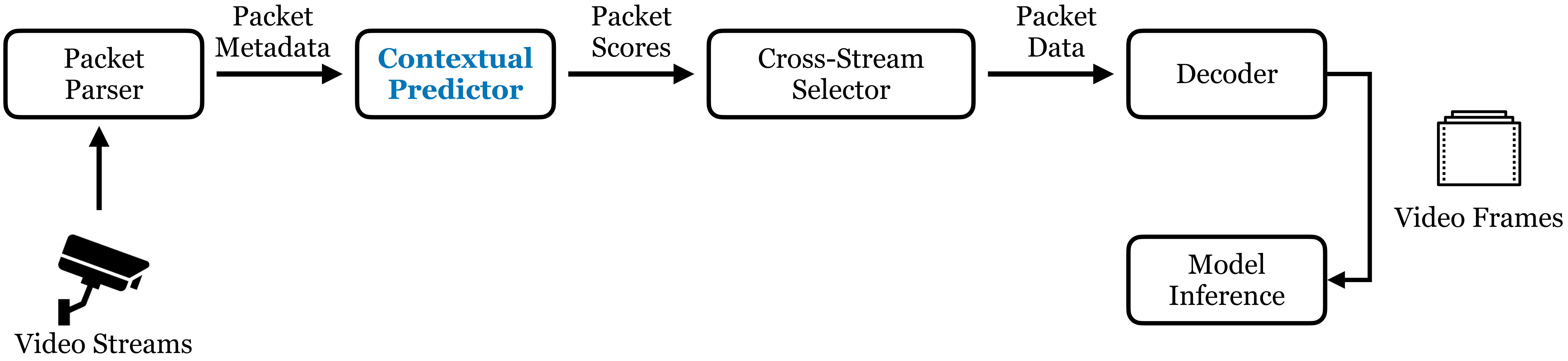
Temporal Estimator

- Available hint#1: historical feedback
  - MAB-based approach



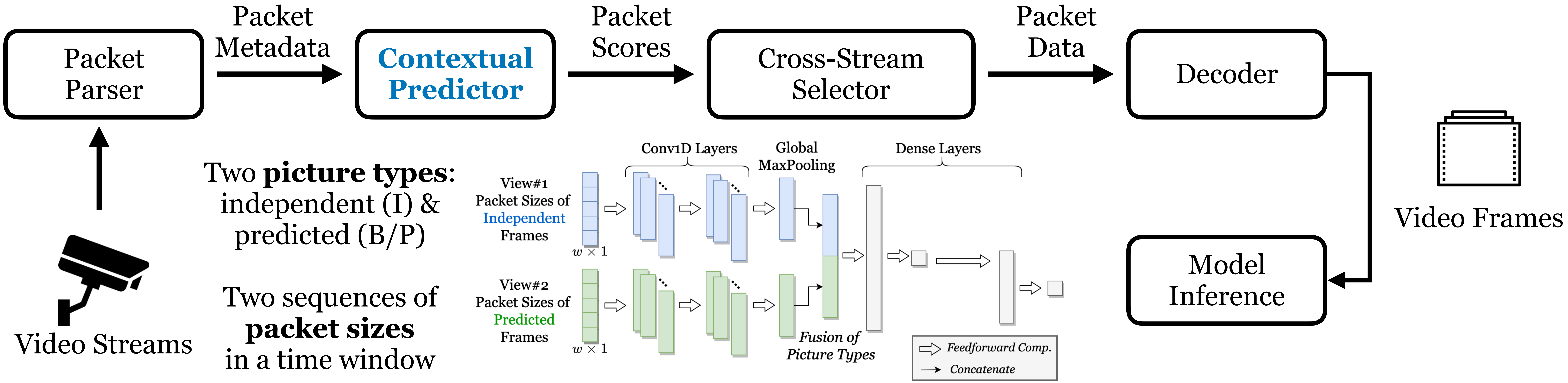
PacketGame Design
Contextual Predictor

- Available hint#2: packet-level metadata
    - Metadata: packet size & picture type



PacketGame Design
Contextual Predictor

- Available hint#2: packet-level metadata
  - Metadata: packet size & picture type
  - Neural network-based predictor

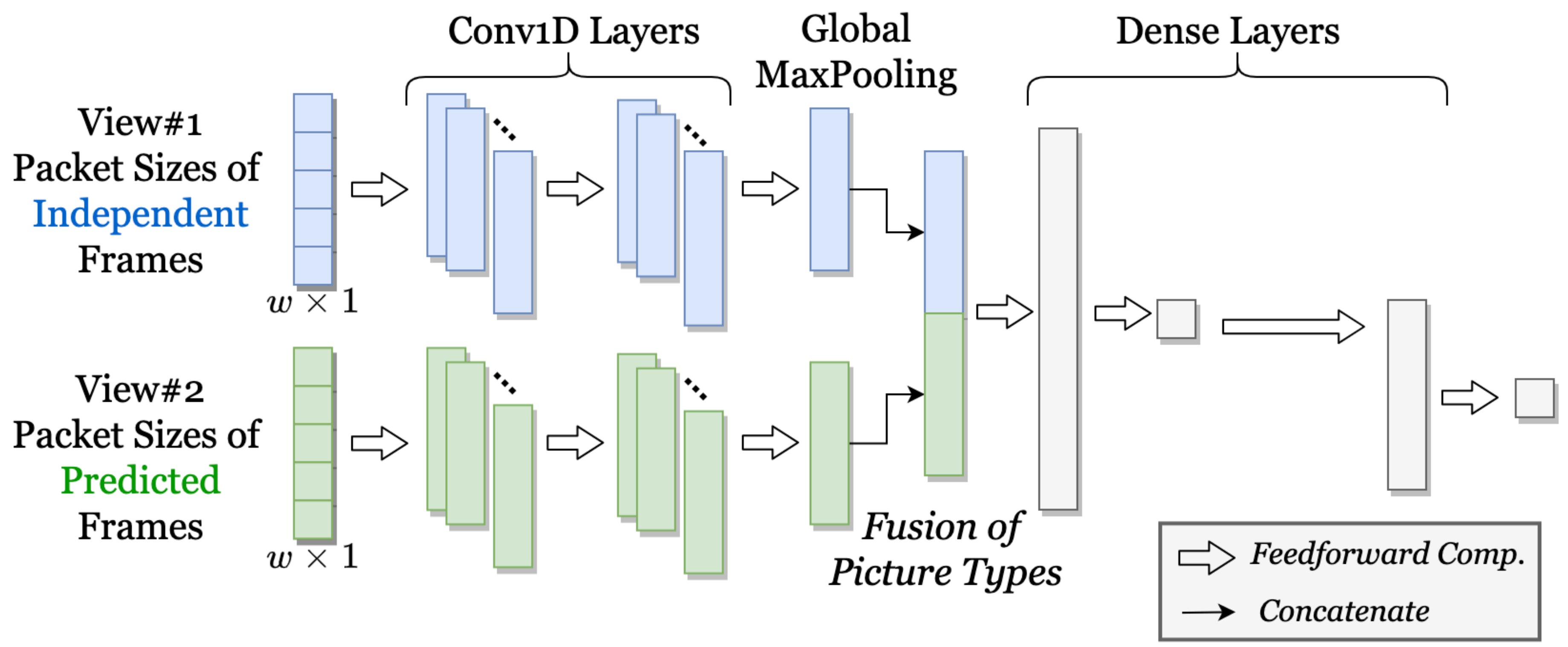


PacketGame Design

Contextual Predictor

- Available hint#2: packet-level metadata
  - Metadata: packet size & picture type
  - Neural network-based predictor
  - Training: offline collected pairs of (X, Y)

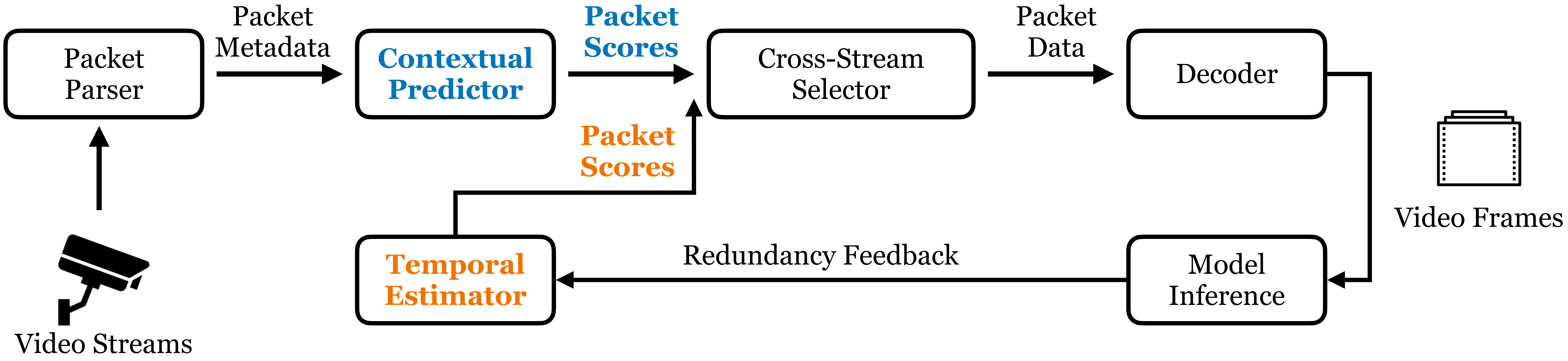
X:
Packet Metadata



Y:
Corresponding Redundancy Label

PacketGame Design
Contextual Predictor

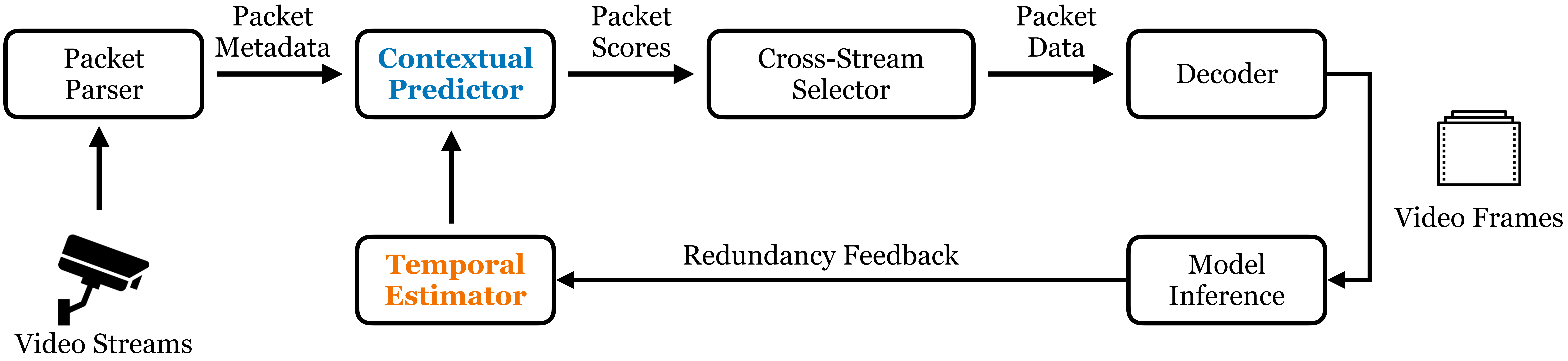
- Packet scores returned by two modules, how to fuse them?



PacketGame Design

Contextual Predictor

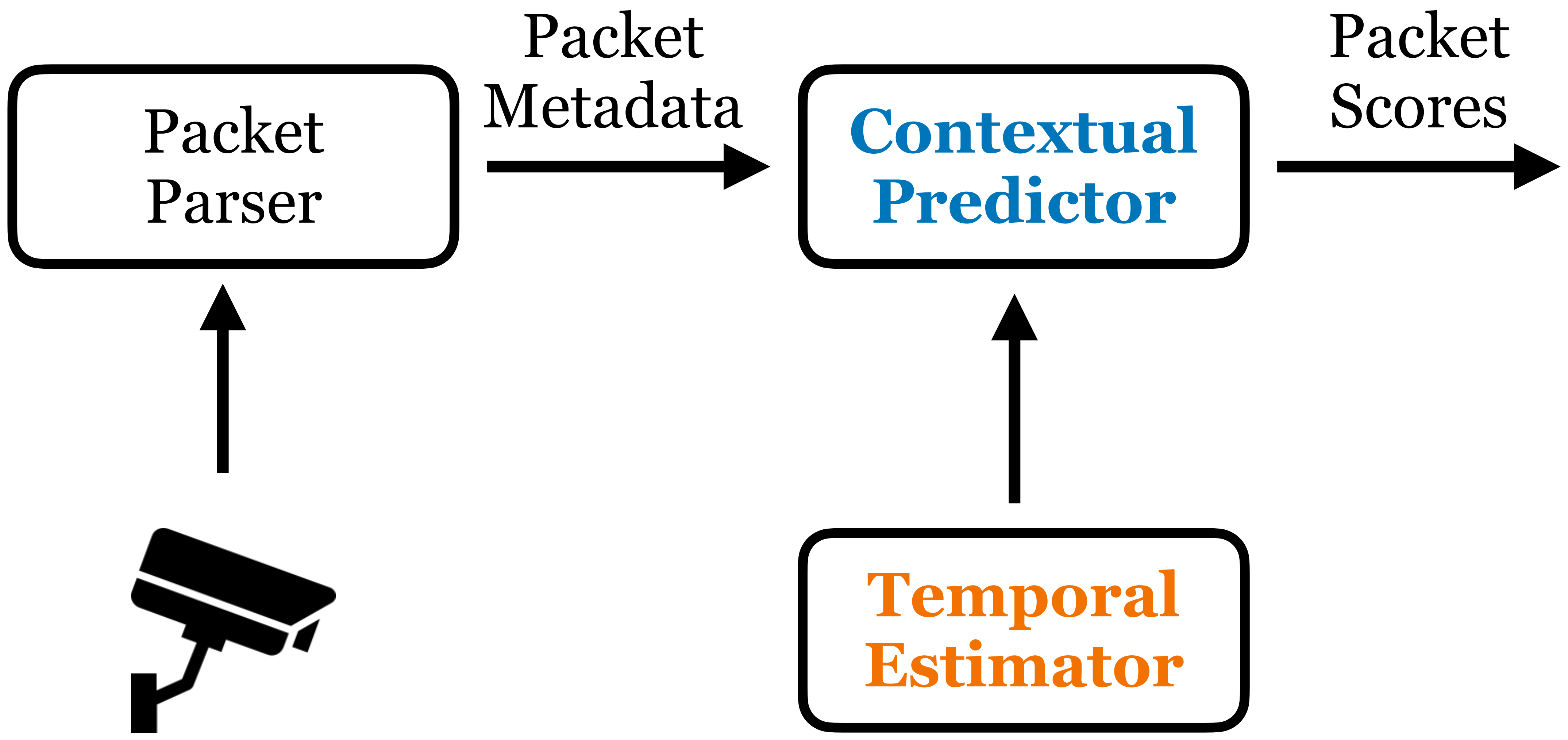
- Fusing the temporal estimator's output as another input view of NN

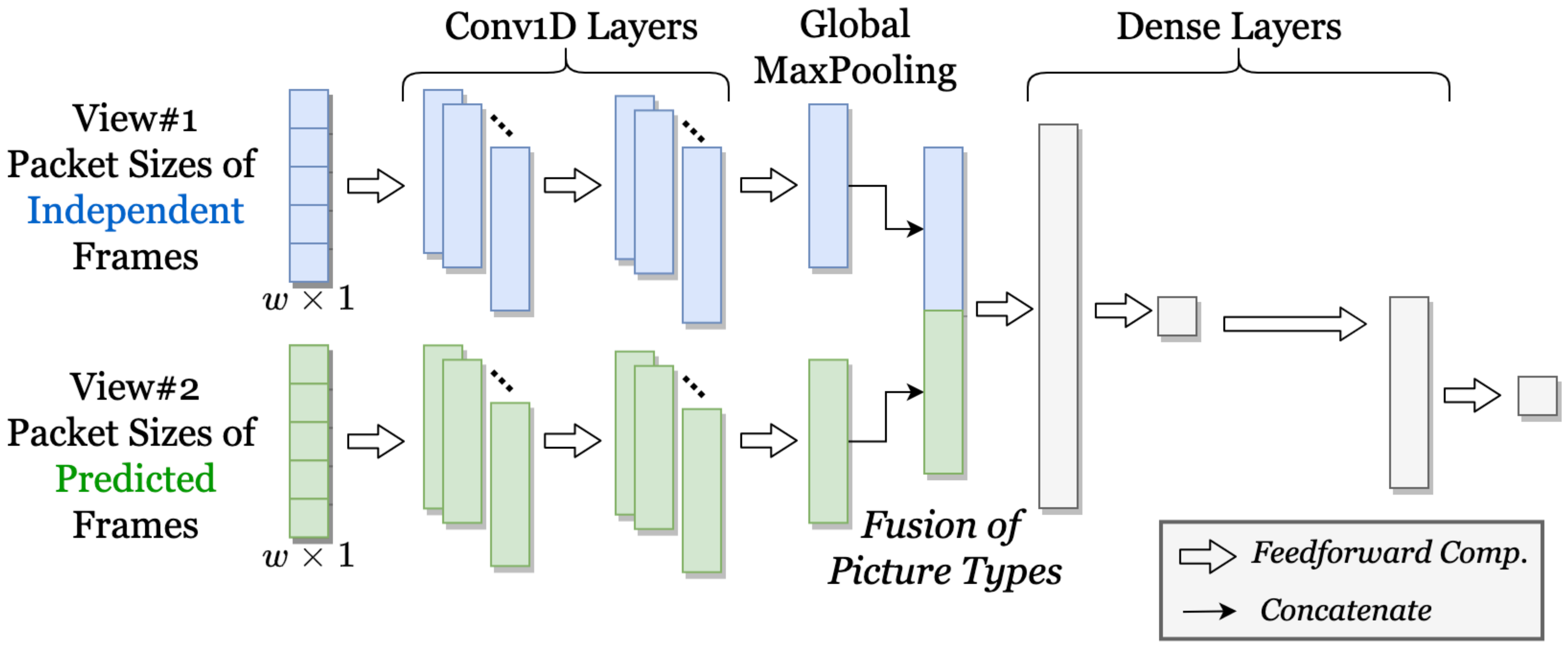


PacketGame Design

Contextual Predictor

- Fusing the temporal estimator's output as another input view of NN

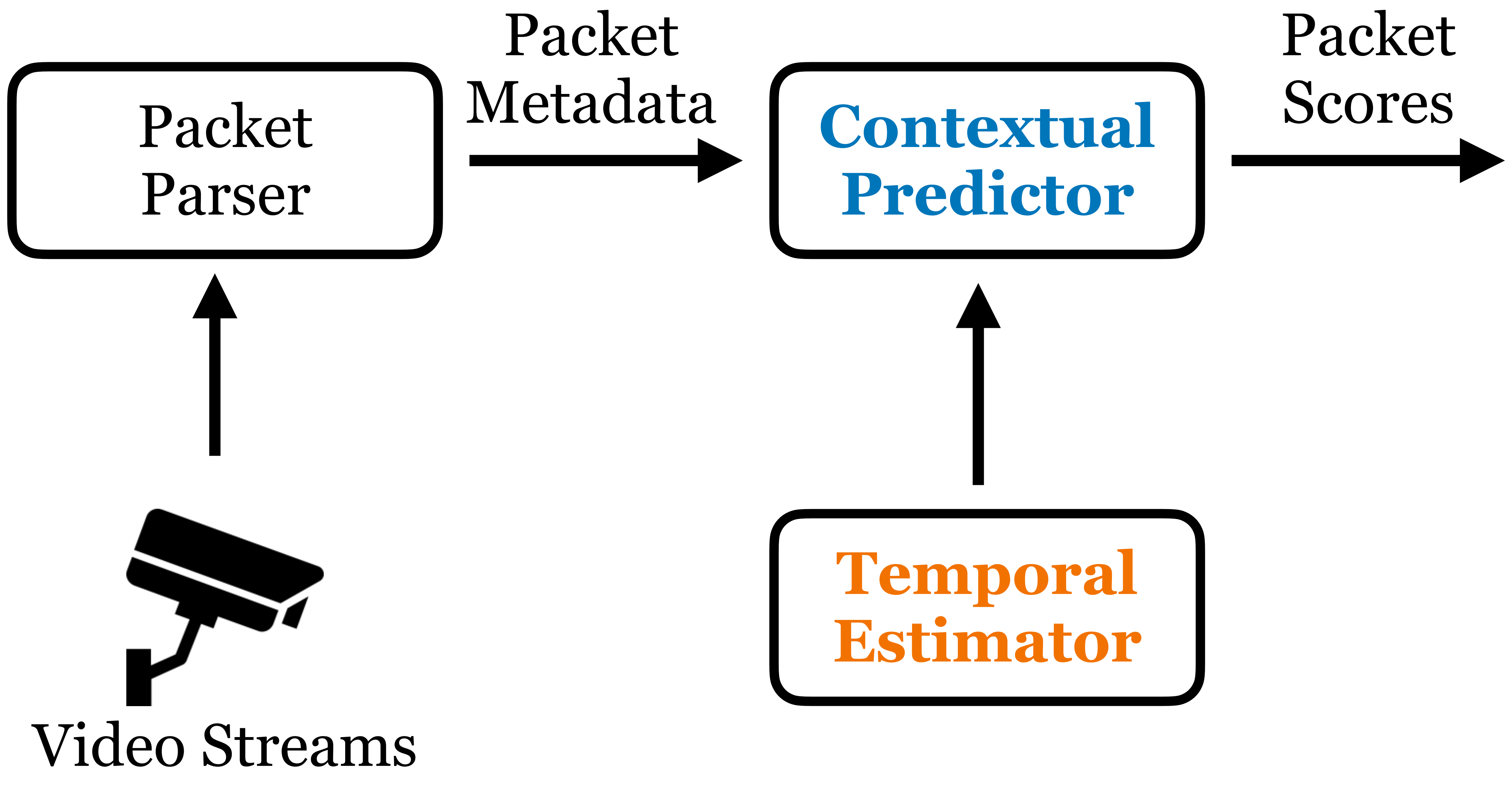


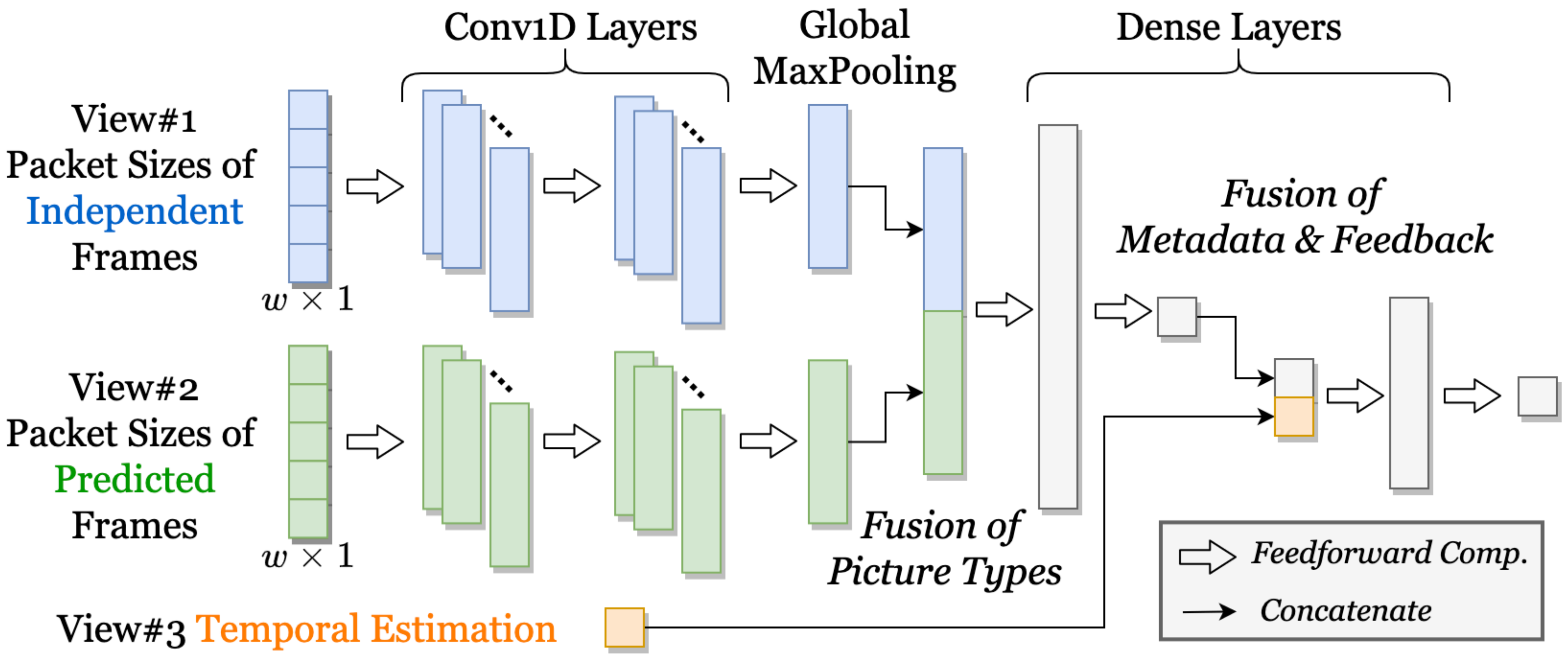


PacketGame Design

Contextual Predictor

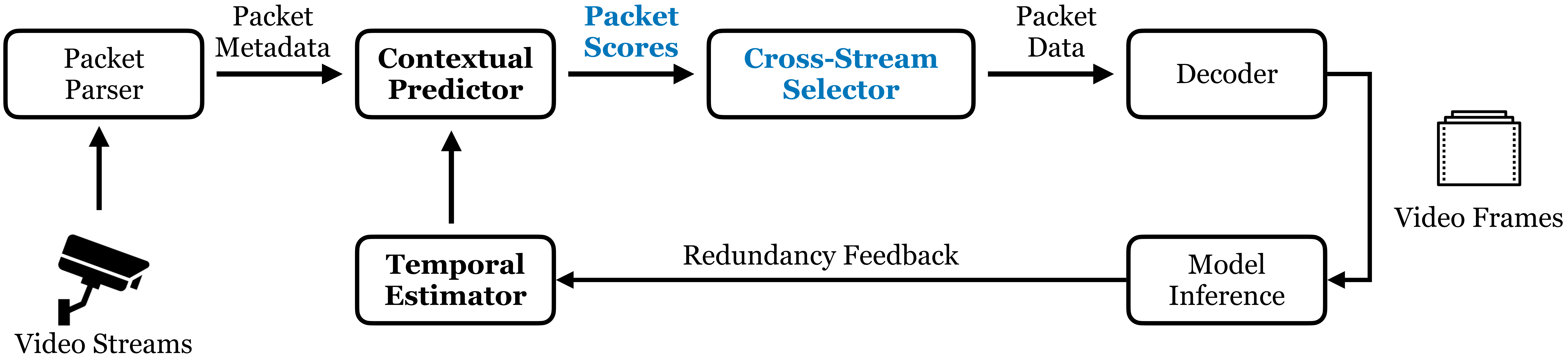
- Fusing the temporal estimator's output as another input view of NN





PacketGame Design
Cross-Stream Selector

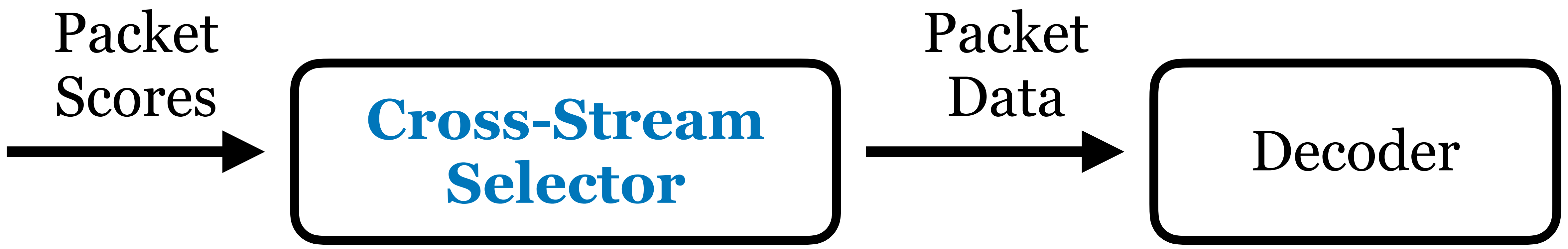
- Combinatorial optimization problem
  - Given predicted packet scores and packet decoding costs, under a decoding budget, maximize the summed scores of selected packets

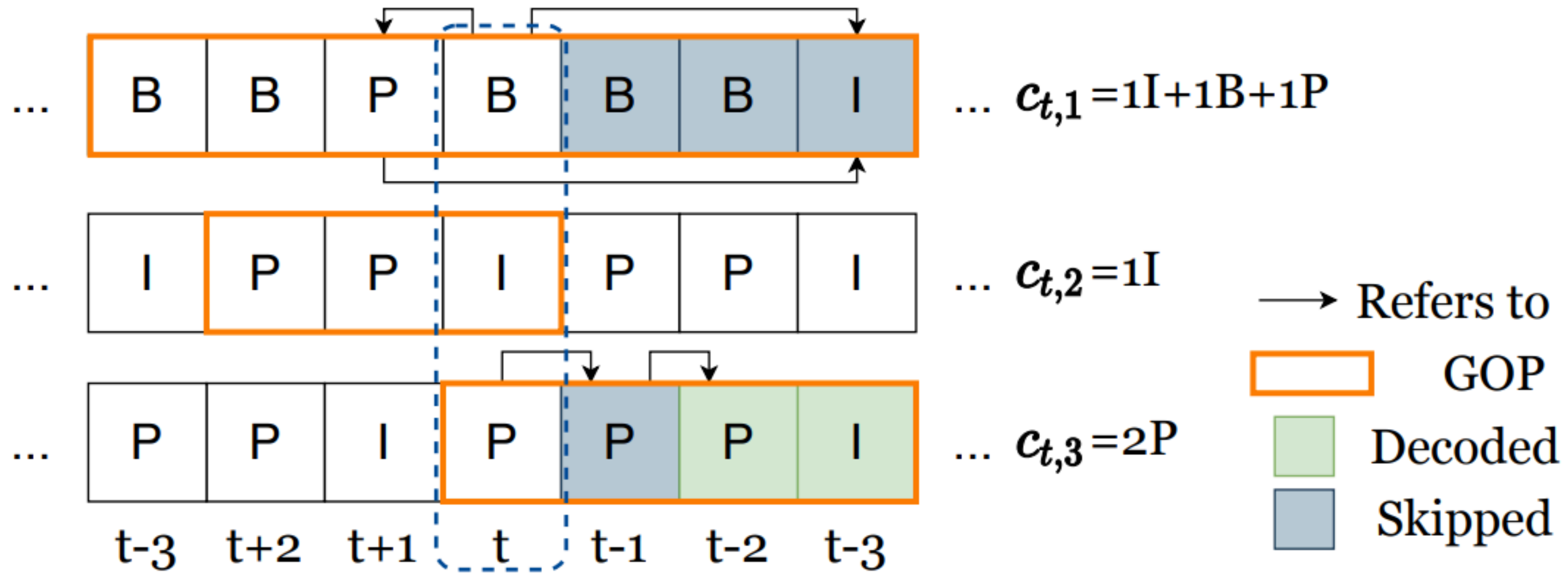


PacketGame Design

Cross-Stream Selector

- Combinatorial optimization problem
  - Given predicted packet scores and packet decoding costs, under a decoding budget, maximize obtained packet scores



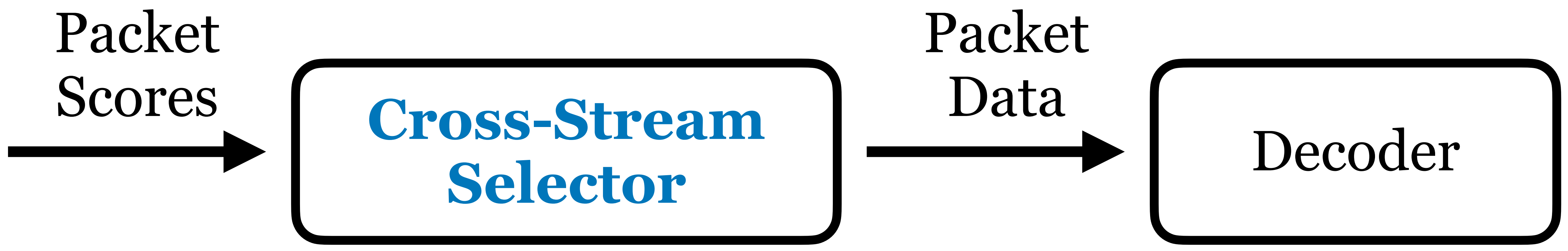


Packet decoding costs depends on picture types and codec configurations (e.g., GOP size)

PacketGame Design

Cross-Stream Selector

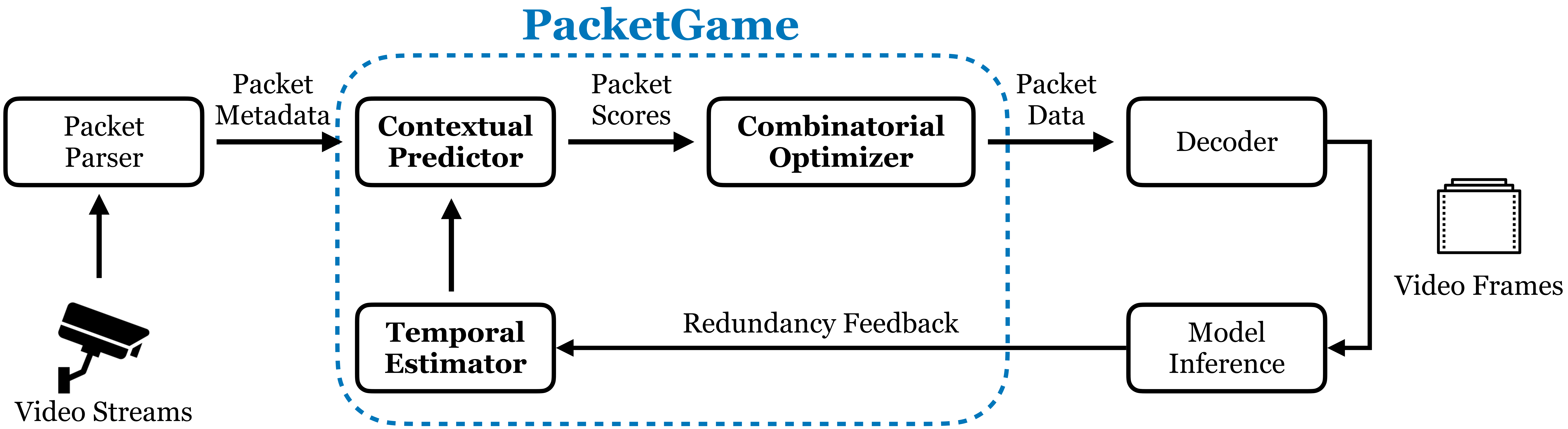
- Combinatorial optimization problem
    - Given predicted packet scores and packet decoding costs, under a decoding budget, maximize obtained packet scores



- we formulate this problem as an approximately fractional knapsack and prove the approximation ratio of the greedy algorithm
    - 1-c/B, in practice, typically greater than 95%

PacketGame Design
Overview

- Overall performance guarantee
  - we prove the regret in T rounds is at most $\tilde{O}(\sqrt{T})$



Outline

- Background
- PacketGame Design
- **Evaluation**

Evaluation
Setting

- 2 public datasets, 1 dataset of collected from campus IP cameras, 3 types of sources
- 4 video inference tasks

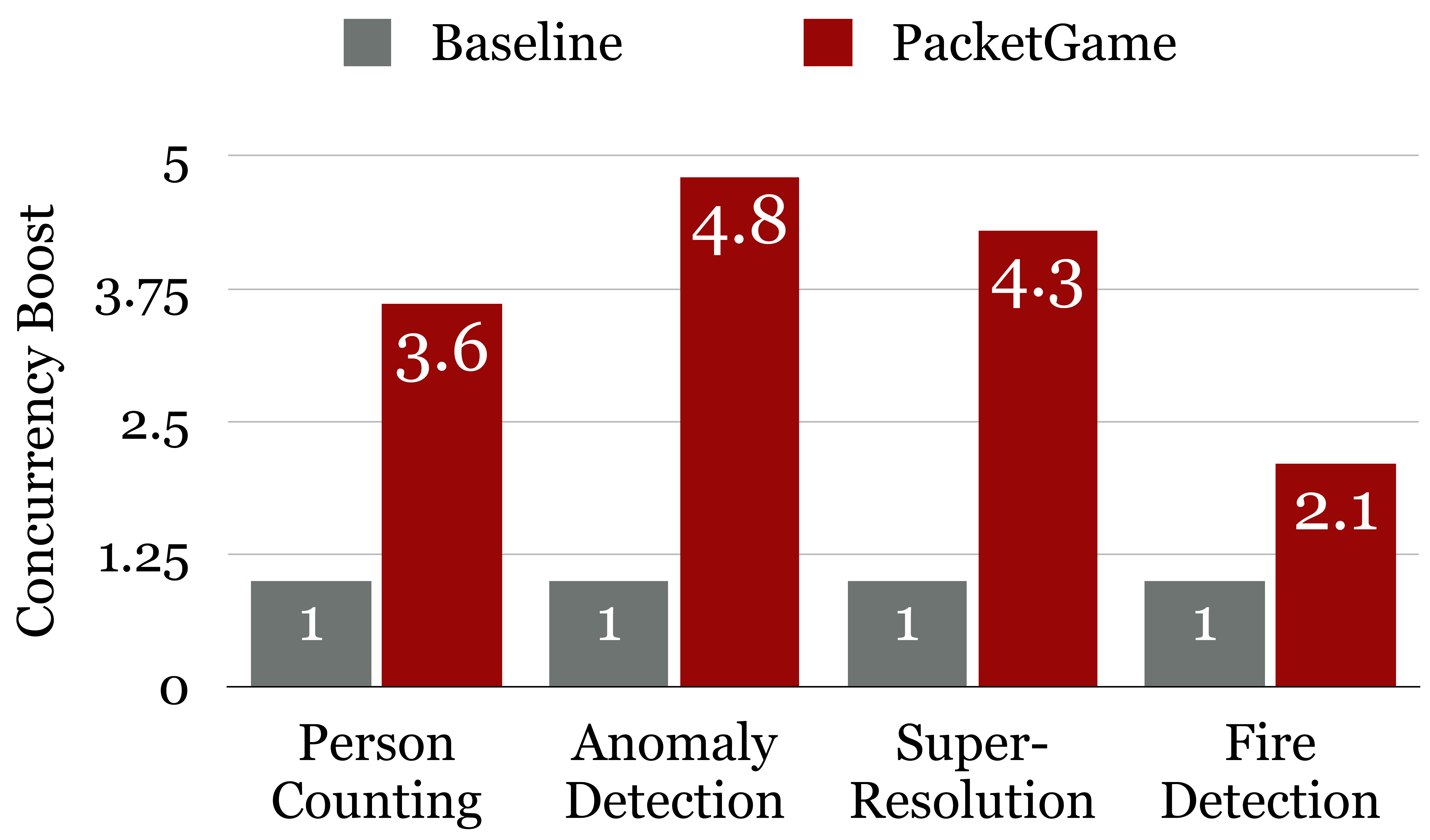
Dataset	Video Source	Inference Task
Campus1K	IP Camera	Person Counting (PC) Anomaly Detection (AD)
YT-UGC	Offline Video	Super-resolution (SR)
FireNet	Mobile Camera	Fire Detection (FD)

- opensource: https://github.com/yuanmu97/PacketGame

Evaluation

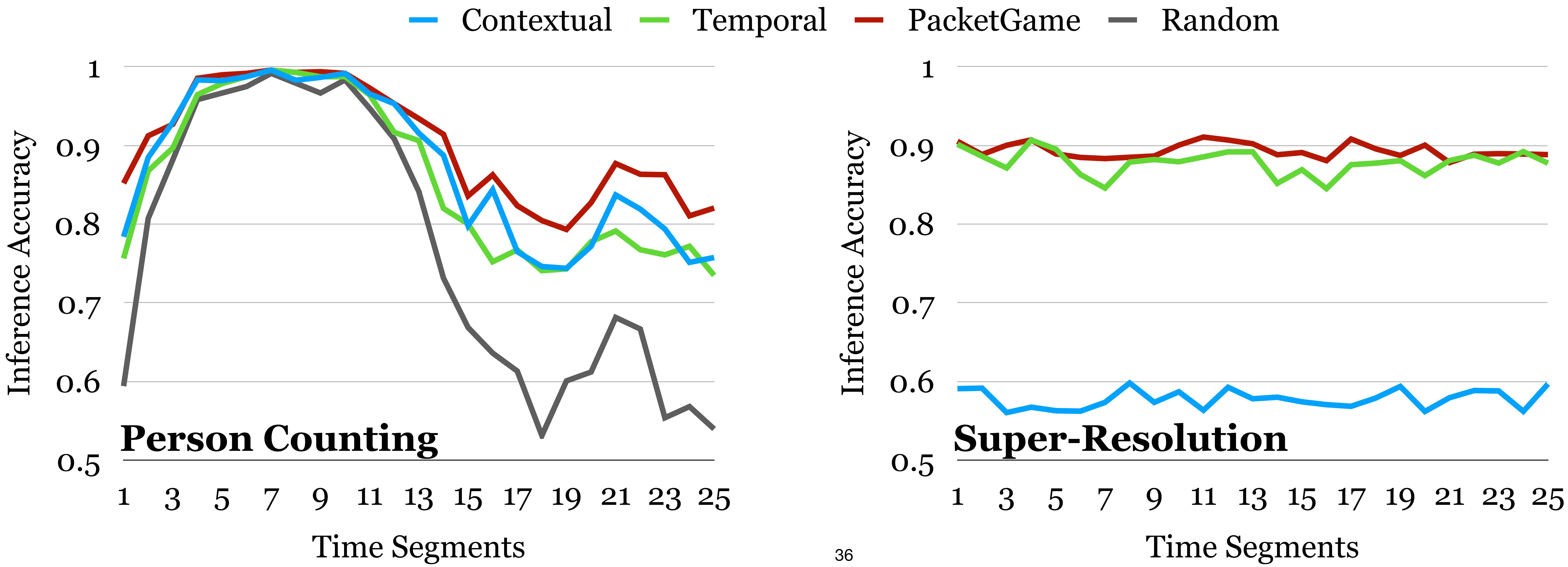
Overall Performance

- target accuracy 90%, PacketGame achieves **2.1-4.8x end-to-end concurrency**



Evaluation
Ablation Study

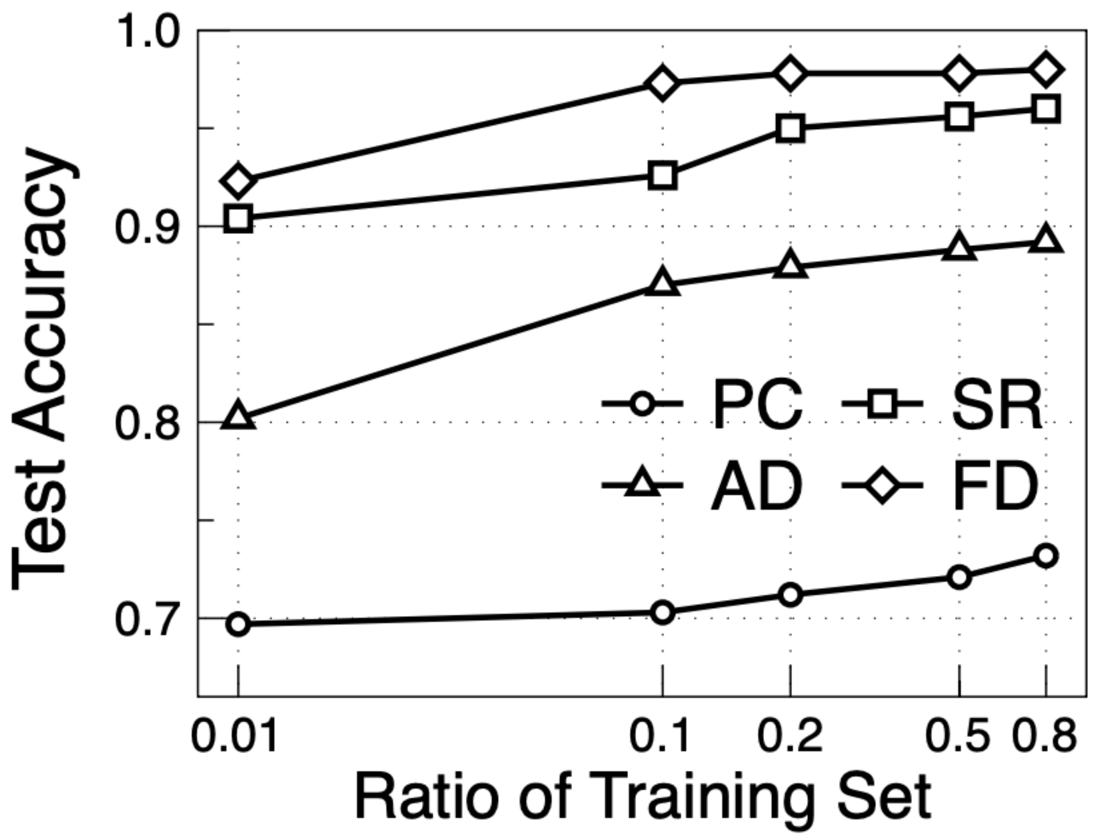
- Contributions of contextual predictor and temporal estimator varies in different tasks



Evaluation

Microbenchmarks

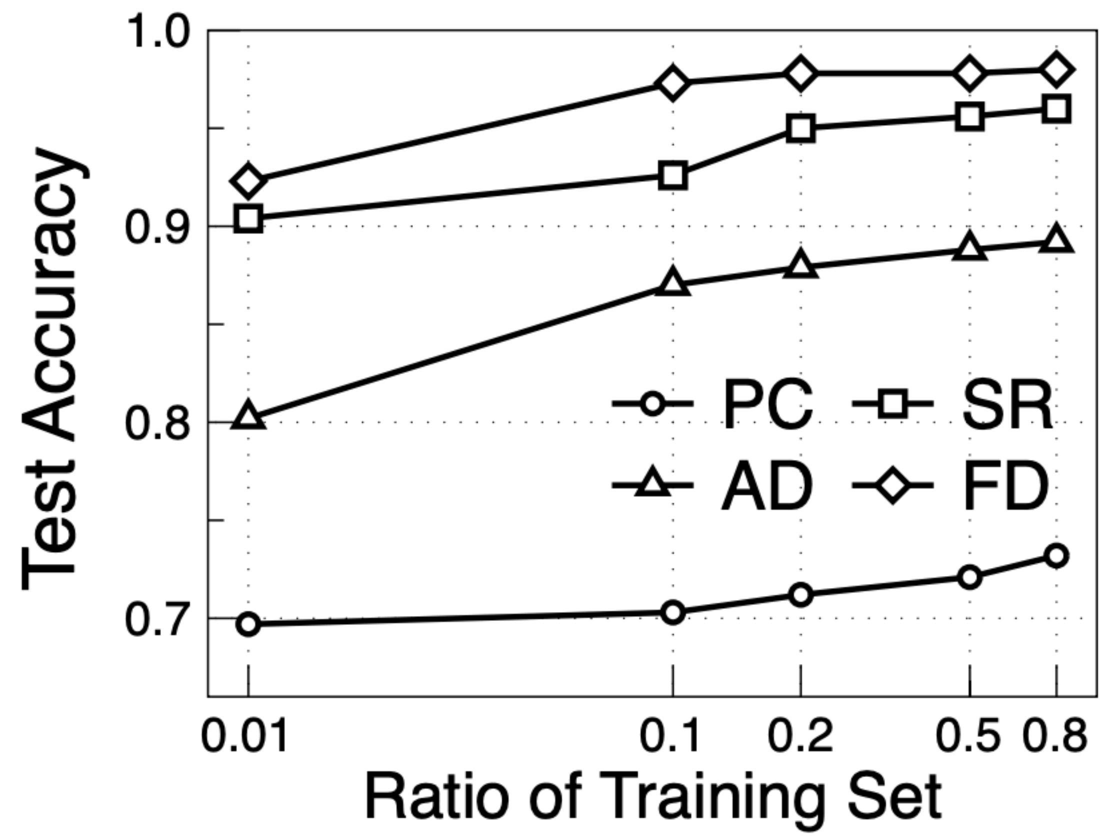
- PacketGame shows robust effectiveness with respect to involved variables, including training size



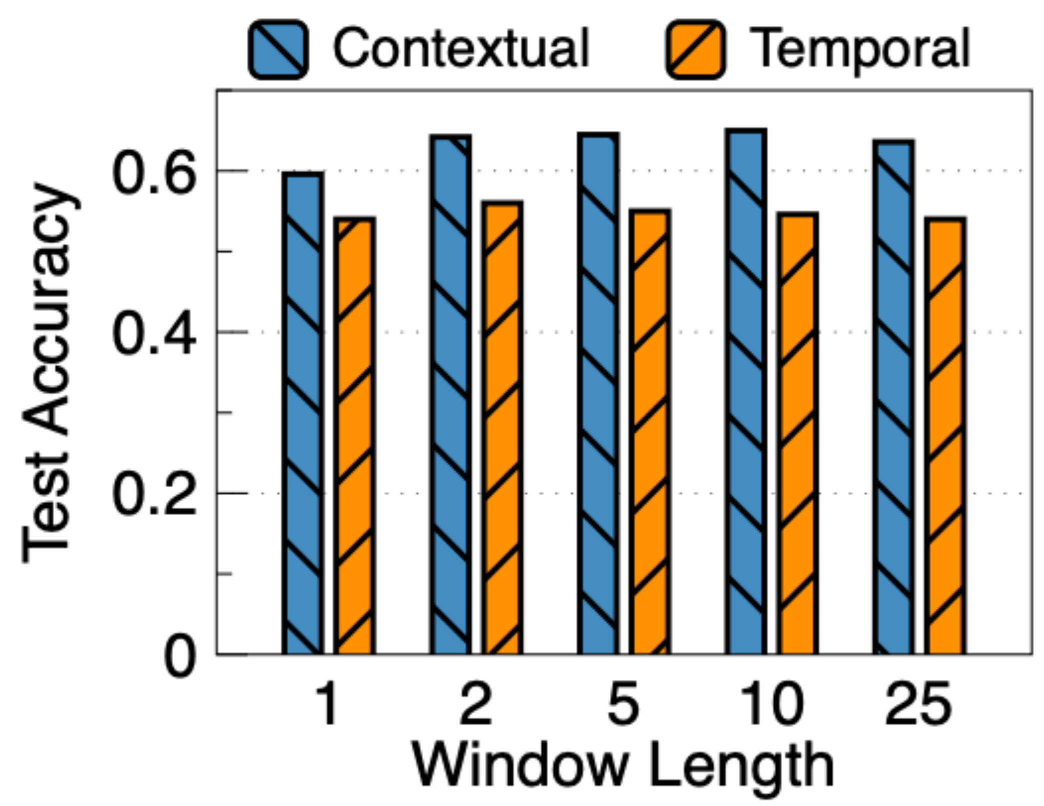
Training Size

Evaluation
Microbenchmarks

- PacketGame shows robust effectiveness with respect to involved variables, including training size, window length



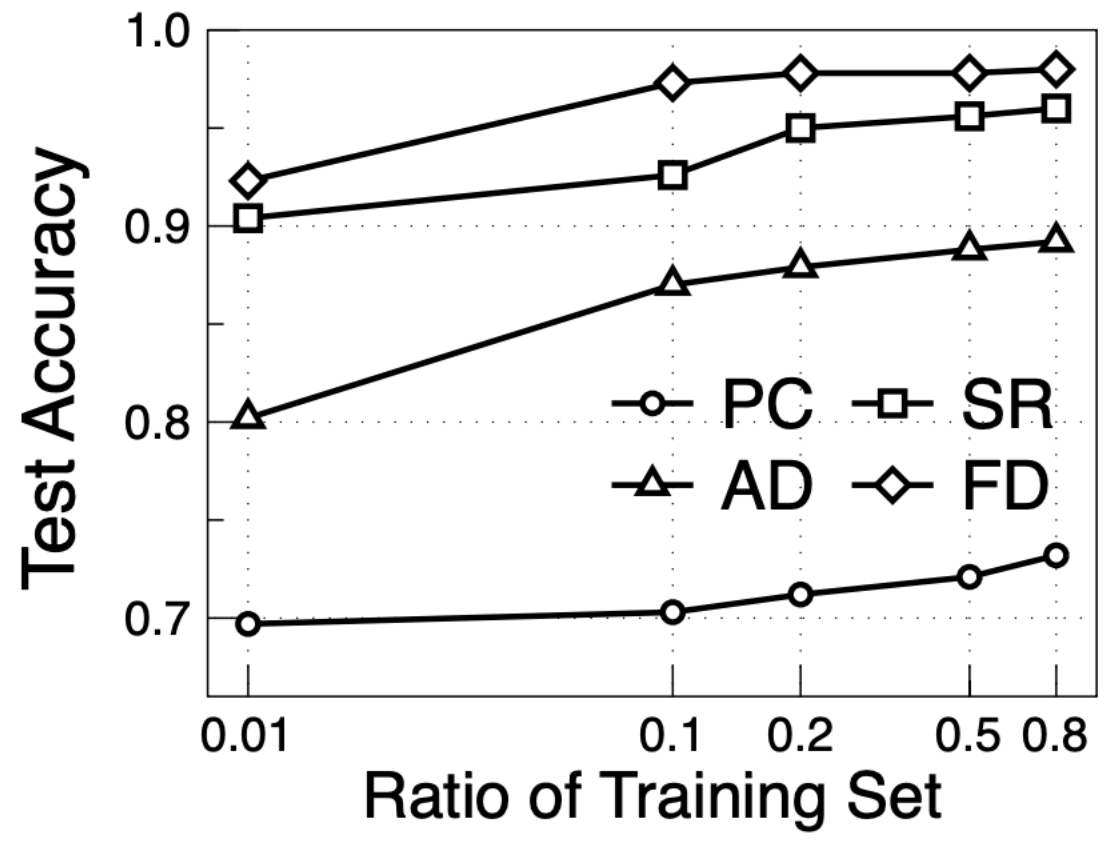
Training Size



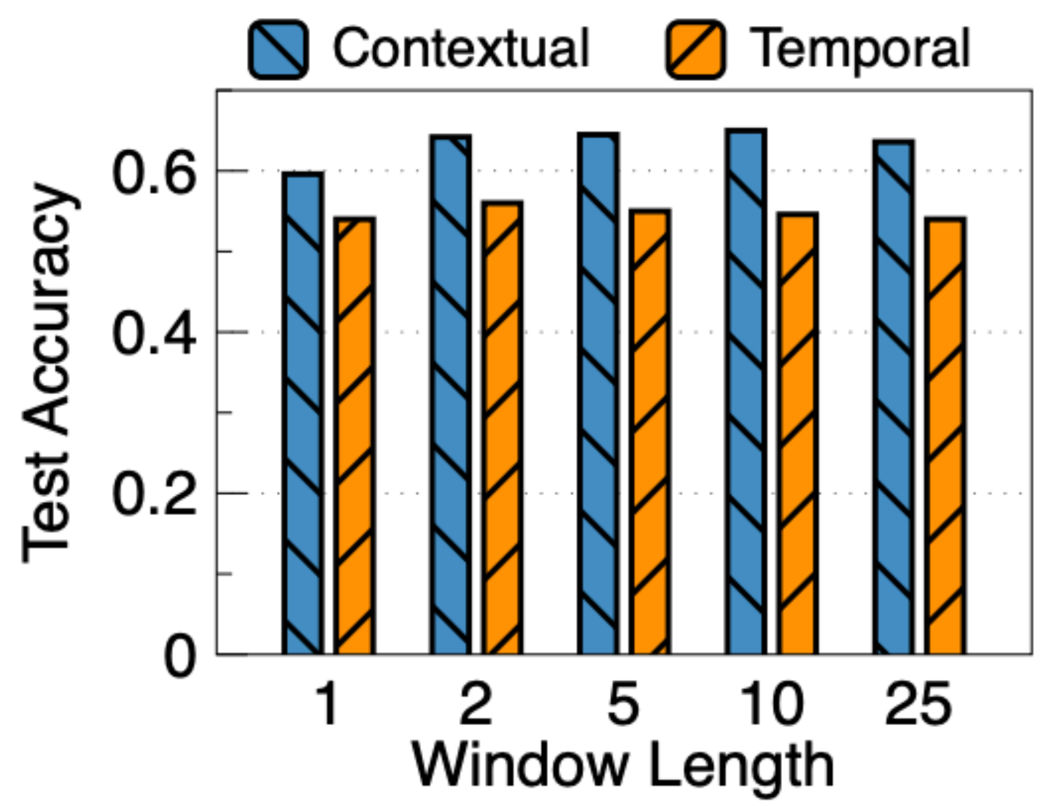
Window Length

Evaluation
Microbenchmarks

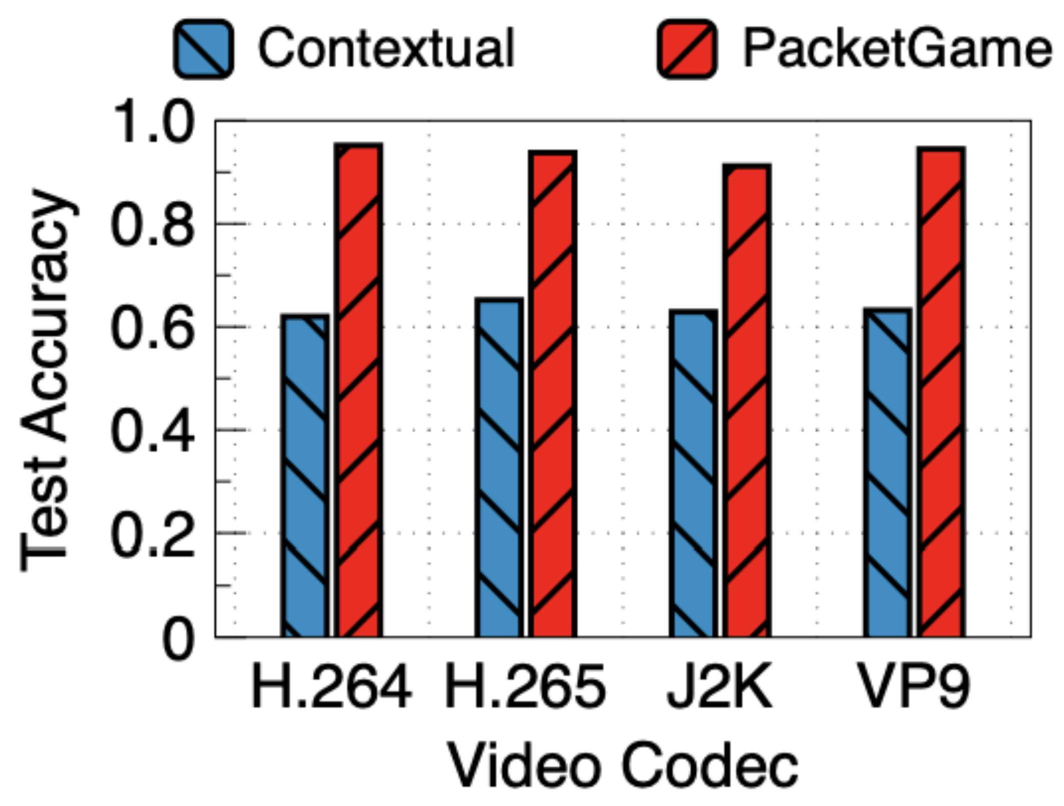
- PacketGame shows robust effectiveness with respect to involved variables, including training size, window length, video codec, etc.



Training Size



Window Length



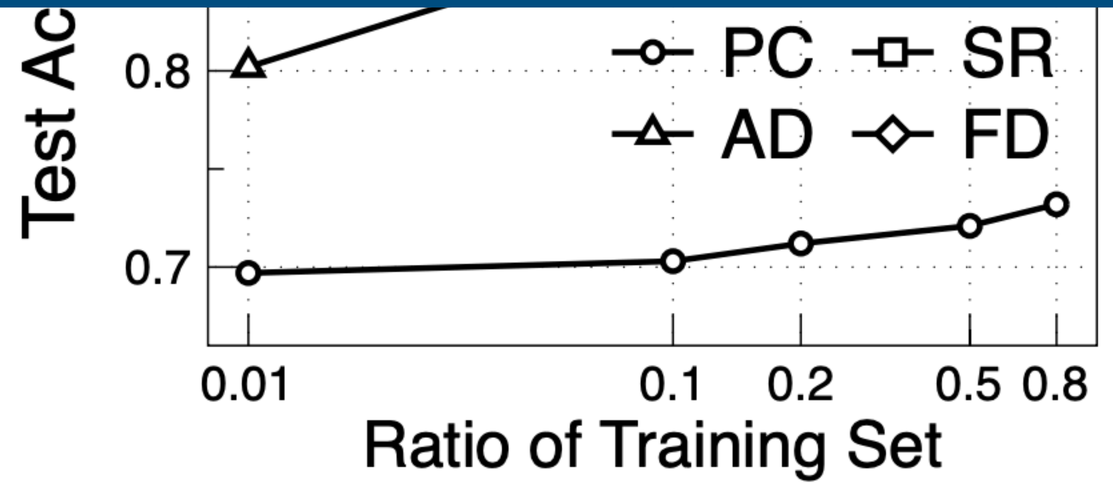
Video Codec

Evaluation

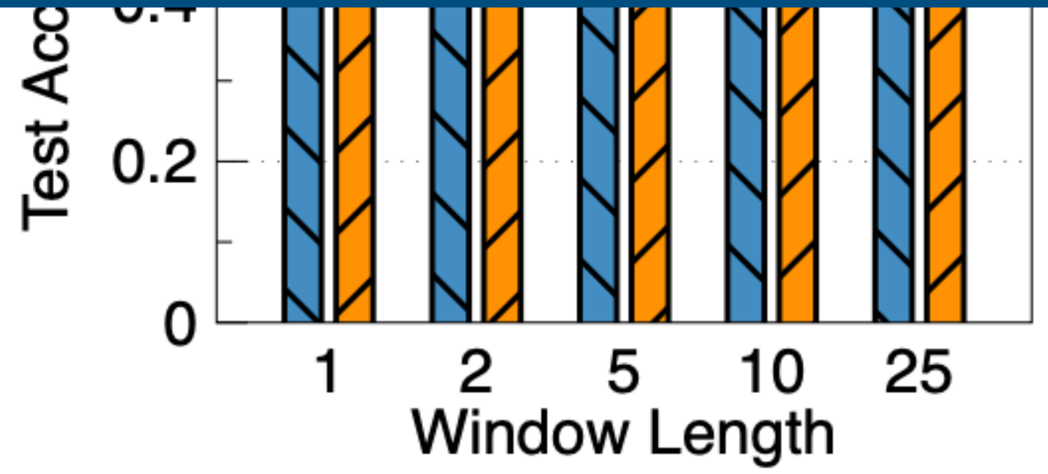
Microbenchmarks

- PacketGame shows robust effectiveness with respect to involved variables, including training size, window length, video codec, etc.

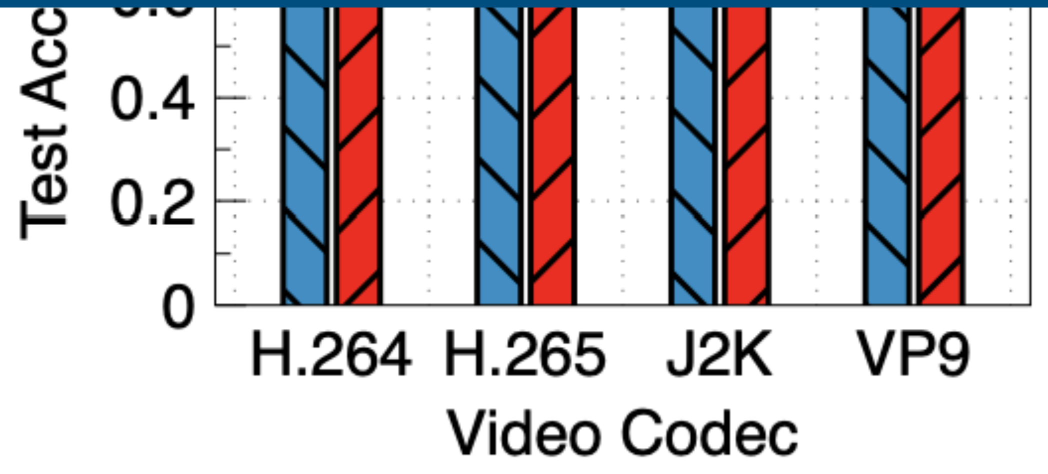
for more about design, theoretical analysis, experimental details, please read our paper :)



Training Size



Window Length



Video Codec

Conclusions
Take-Home Messages

- The system bottleneck for multi-module pipeline is constantly changing, and now it's the **decoder's** turn for large-scale video analytics.
- Packet gating is promising and easy to implement. Try PacketGame for your video analytics system :)
- In the future, similar ideas could be explored for packet-level selection of other modalities, like audio and motion signals. Hope to inspire your research!

Acknowledgement

- My advisors Prof. Xiang-Yang Li and Prof. Lan Zhang in LINKE lab.

- Researchers and engineers at IAI 合肥综合性国家科学中心人工智能研究院 Institute of Artificial Intelligence, Hefei Comprehensive National Science Center for developing our video analytics system at USTC.
    - Xuanke You, Shanyang Jiang, Miao-Hui Song, Changhu Can, Yueting Liu, Qing Chu, Ke Ding, Jin Yan

Thanks!
Q&A

Mu Yuan
Email: ym0813@mail.ustc.edu.cn