Titan: A Two-Stage Data Selection Framework for Data-Efficient Model Training on Edge Devices

Chen Gong, Rui Xing, Zhaenzhe Zheng, Fan Wu
Shanghai Jiao Tong University

2025-08-07

August 3-7, 2025
KDD2025

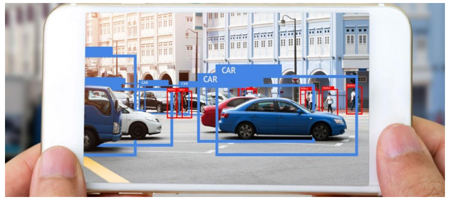
Outline

1. **Background**
2. Challenges
3. Titan Design
4. Evaluation

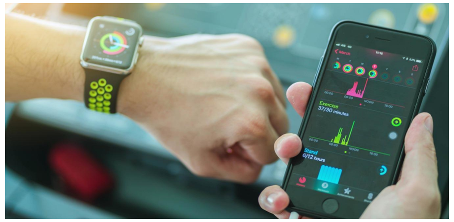
On-Device ML

ML models are widely deployed across various mobile apps.

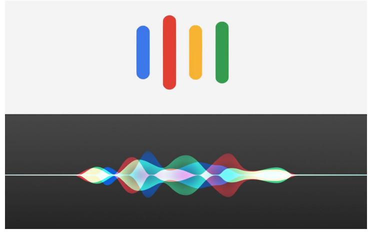
Image Analysis



Human Activity Recognition

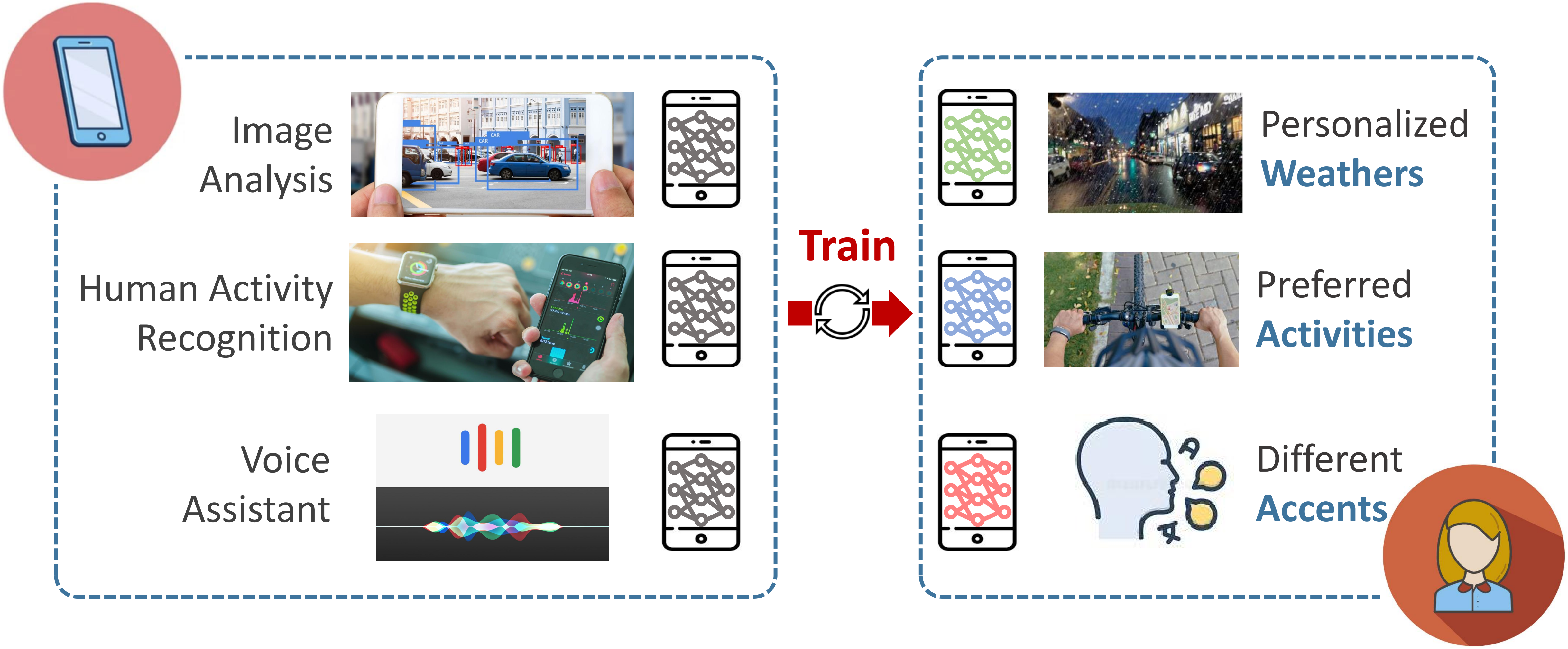


Voice Assistant



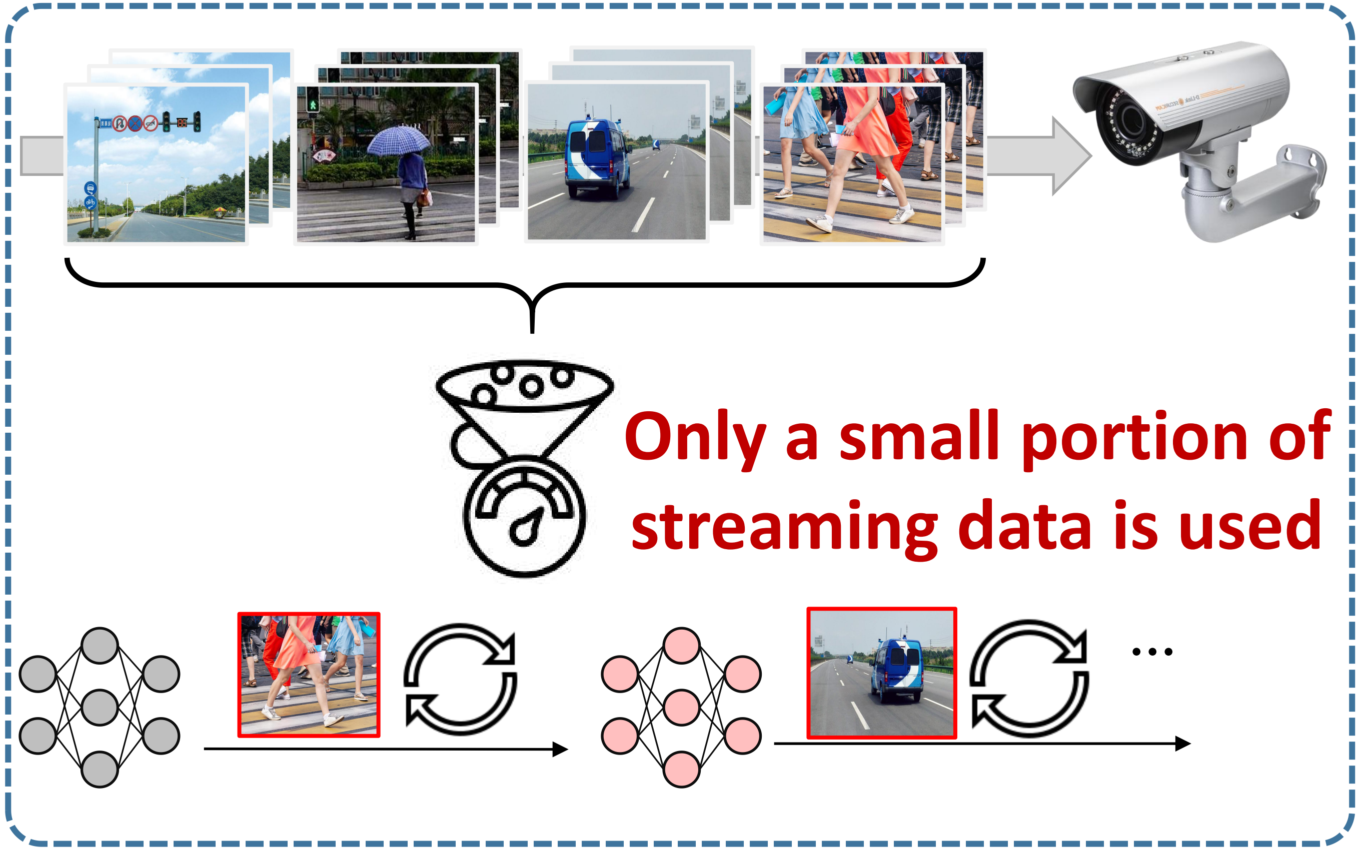
On-Device ML

On-device model training is critical for personalization & privacy.



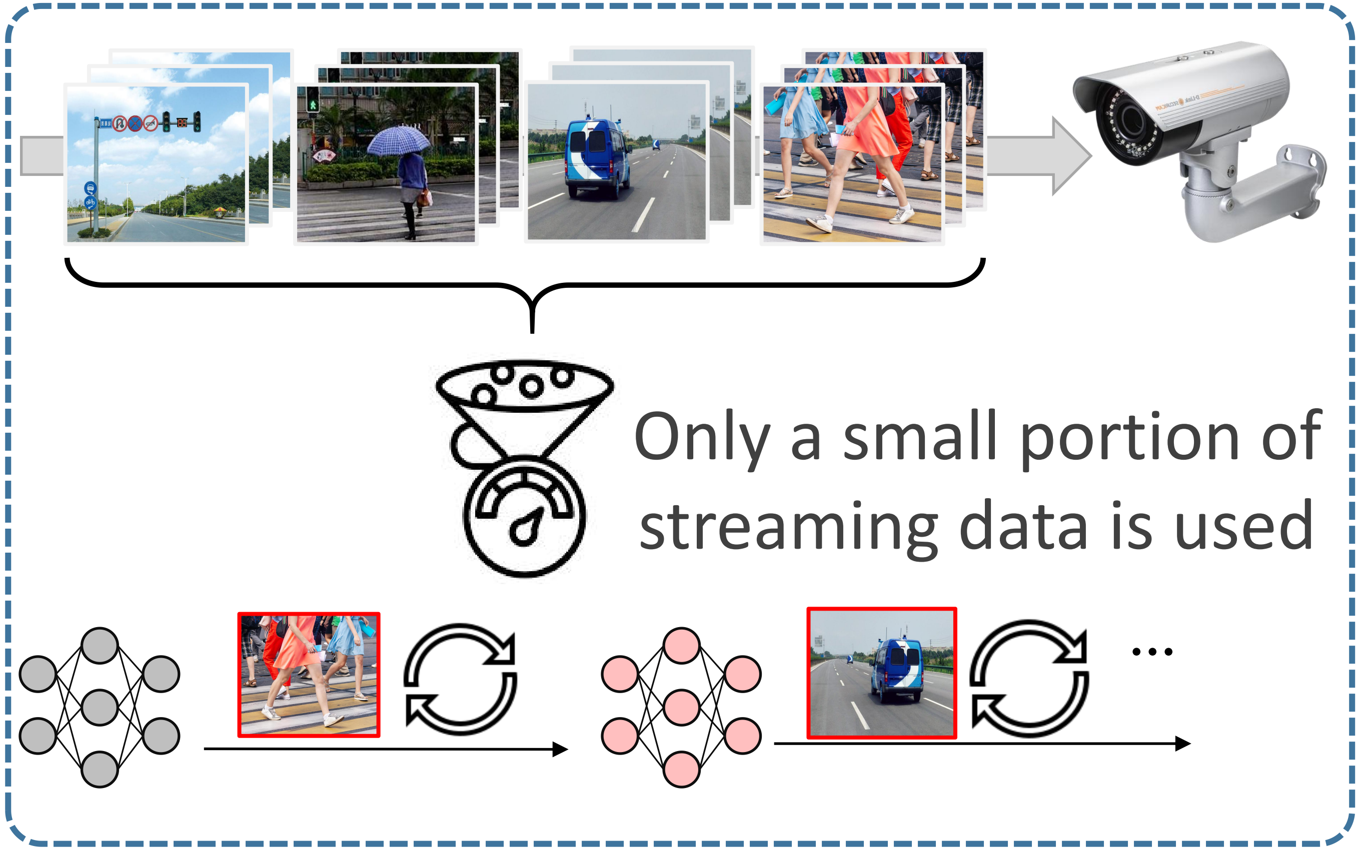
Data Bottleneck

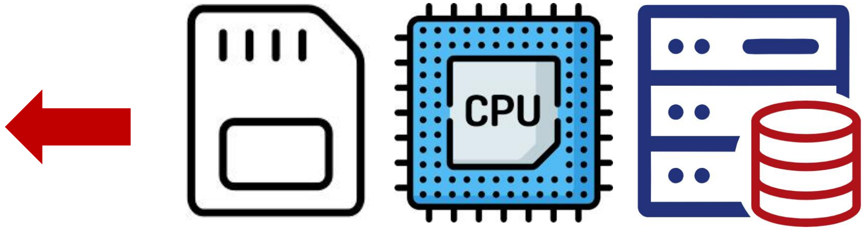
Under-utilization of on-device data stream is a key bottleneck.



Data Bottleneck (Cont.)

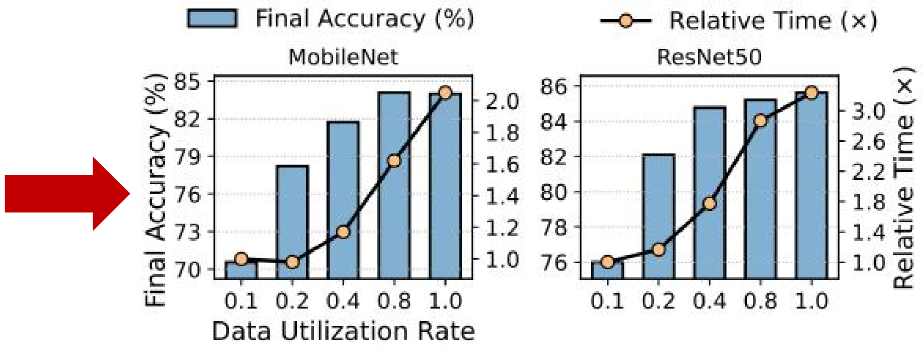
Under-utilization of on-device data stream is a key bottleneck.





Limited Device Resources

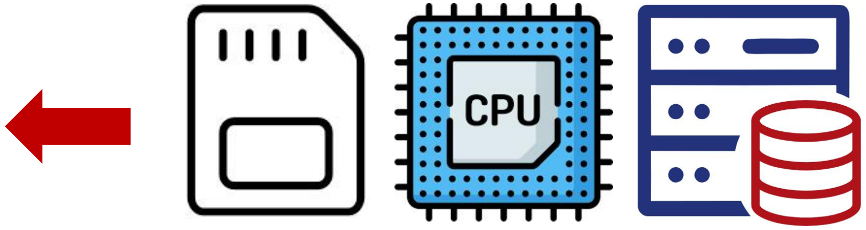
(Memory, Computation, Storage)



Acc: 13%↓ Time: 3.1x↑

Data Bottleneck (Cont.)

Under-utilization of on-device data stream is a key bottleneck.



Key Problem: Improve on-device model training performance by prioritizing important data

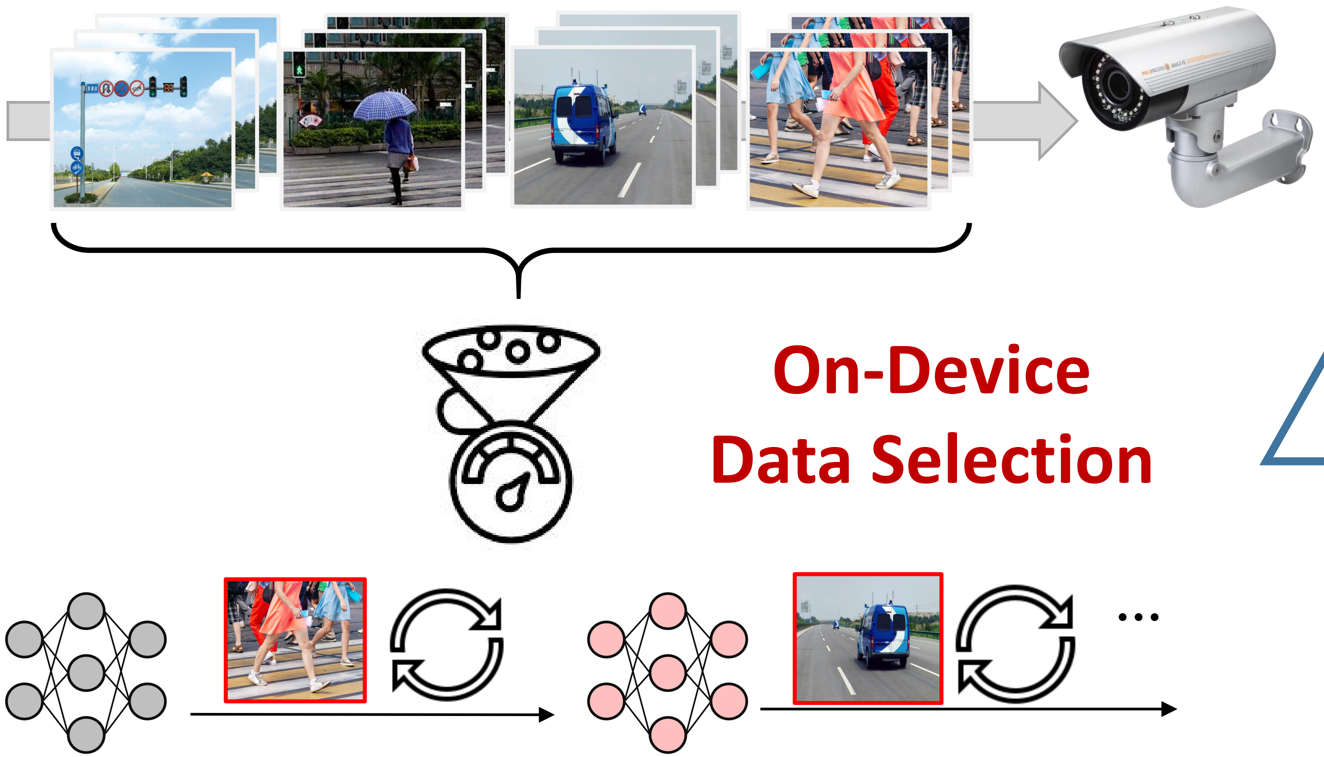
Acc: 13%↓ **Time: 3.1x↑**

Outline

1. Background
2. **Challenges**
3. Titan Design
4. Evaluation

Design Challenges

Achieving effectiveness and efficiency is critical but challenging.

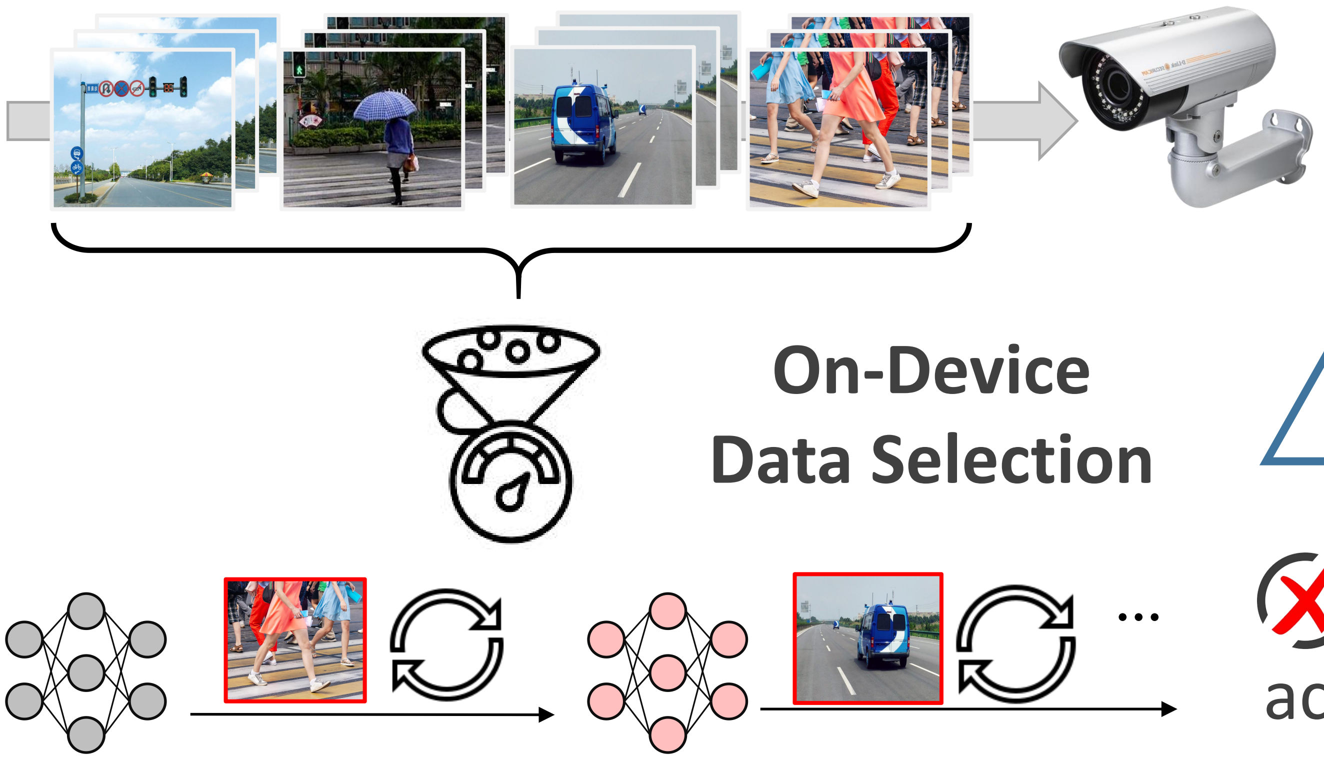


Effectiveness:
theoretical & empirical guarantees.

Efficiency:
time & resource efficiency.

Design Challenges

Achieving effectiveness and efficiency is critical but challenging.



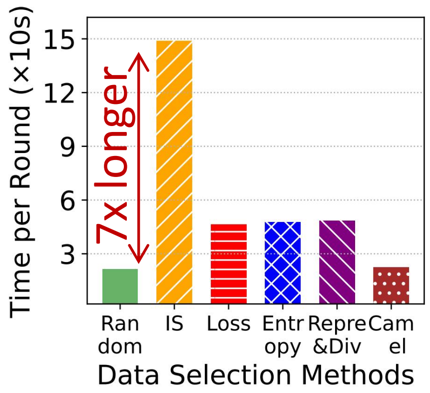
Effectiveness:
theoretical & empirical guarantees.
Efficiency:
time & resource efficiency.

Higher effectiveness →
accurate but costly evaluations on more data → **low efficiency**

Limitations of Existing Works

Cloud-side data selection methods fail to work for device side.

- **Importance Sampling** (*IS*): select data according to *gradient norms*.

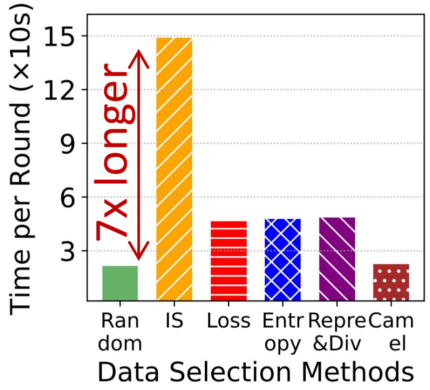


High evaluation latency for data stream

Limitations of Existing Works

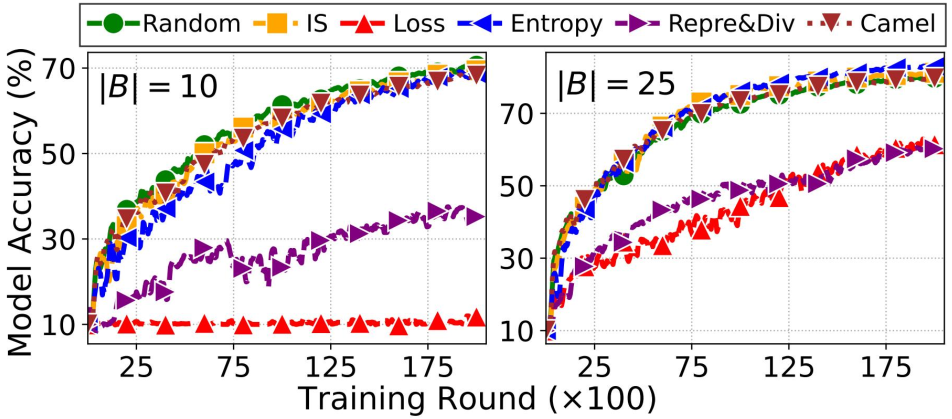
Cloud-side data selection methods fail to work for device side.

- **Importance Sampling** (*IS*): select data according to *gradient norms*.



High evaluation latency for data stream

- **Heuristic Selection**: prioritize data with high *loss, entroy, rep&div.*
- **Coreset Selection** (*Camel*): choose a weighted data-subset with highest *gradient similarity.*



Low performance gains w.r.t. random sampling

Outline

1 Background

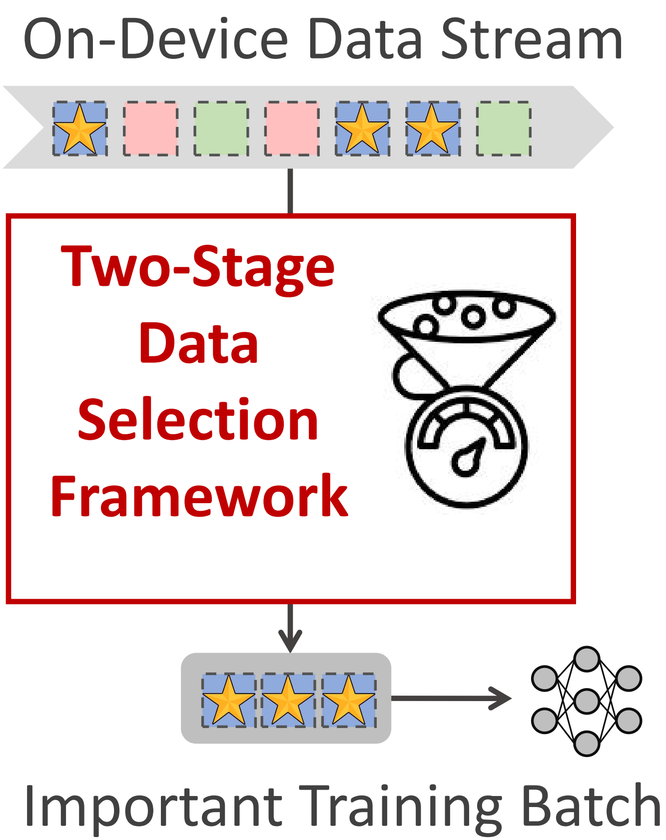
2 Challenges

3 **Titan Design**

4 Evaluation

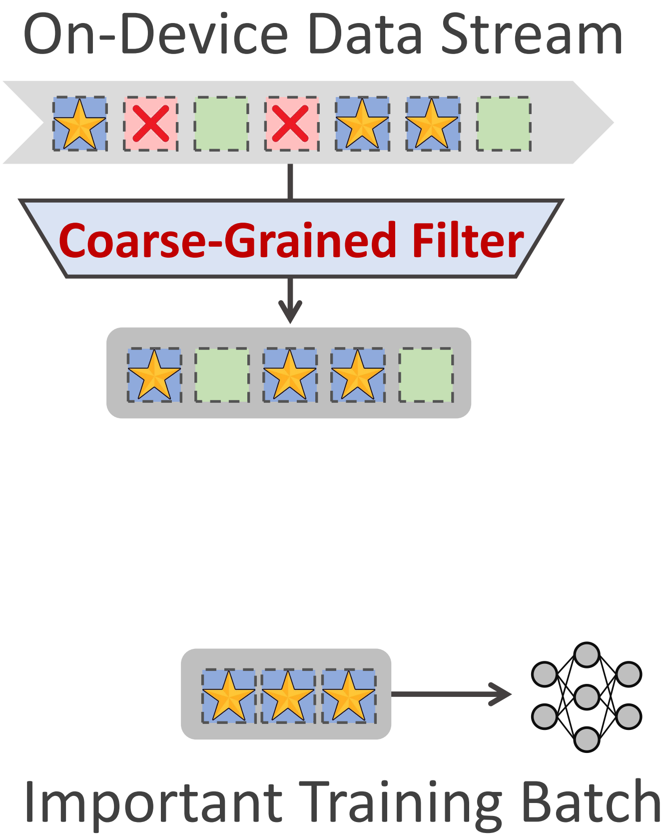
Overall Design

Goal: Exploit on-device data resources efficiently and effectively.



Overall Design (Cont.)

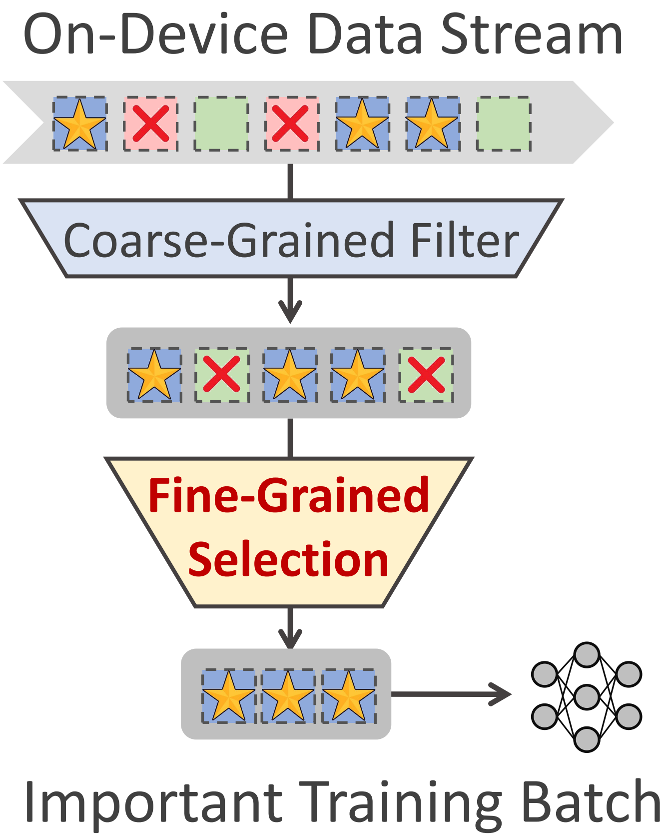
Goal: Exploit on-device data resources efficiently and effectively.



- **Time-Efficiency:** A coarse-grained filter to filter out a small candidate dataset.

Overall Design (Cont.)

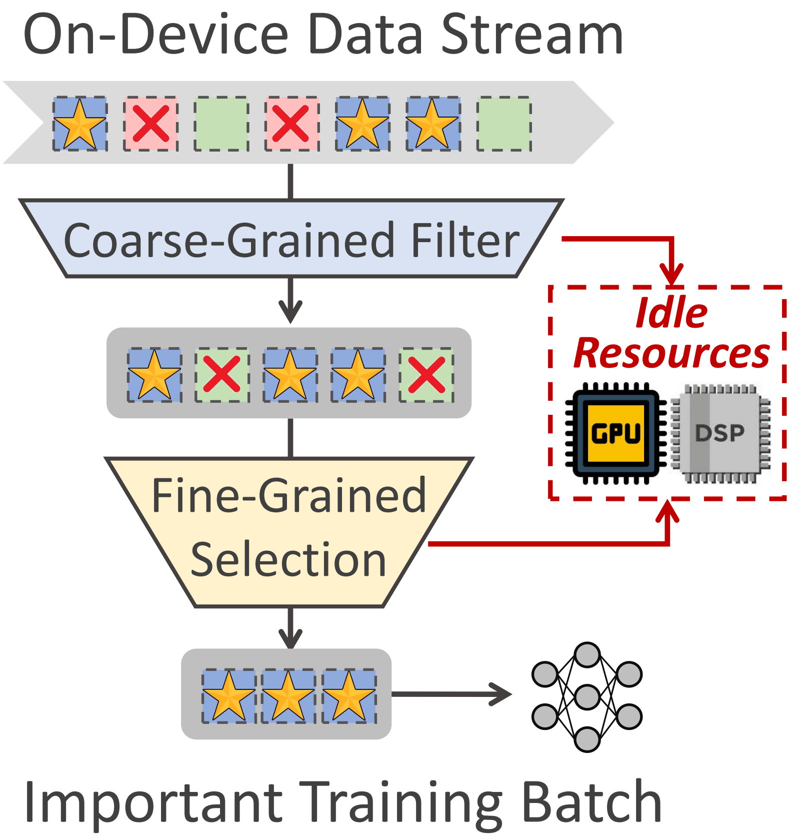
Goal: Exploit on-device data resources efficiently and effectively.



- Time-Efficiency: A coarse-grained filter to filter out a small candidate dataset.
- **Effectiveness:** A theoretically optimal data selection algorithm to maximize performance.

Overall Design (Cont.)

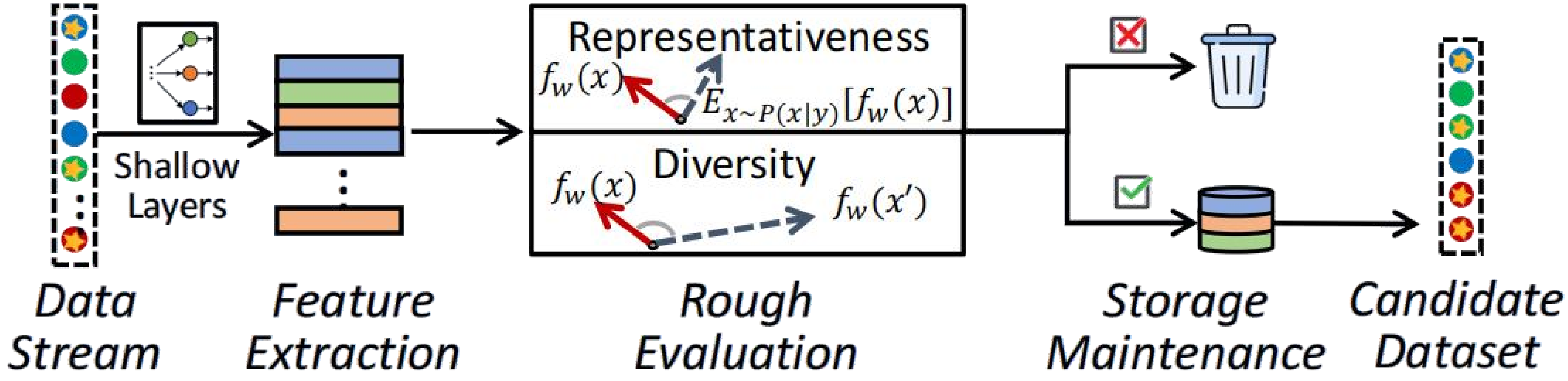
Goal: Exploit on-device data resources efficiently and effectively.



- Time-Efficiency: A coarse-grained filter to filter out a small candidate dataset.
- Effectiveness: A theoretically optimal data selection algorithm to maximize performance.
- **Resource-Efficiency:** A pipeline design to offload data selection to idle resources.

Coarse-Grained Filter

Filter streaming data via representativeness and diversity.



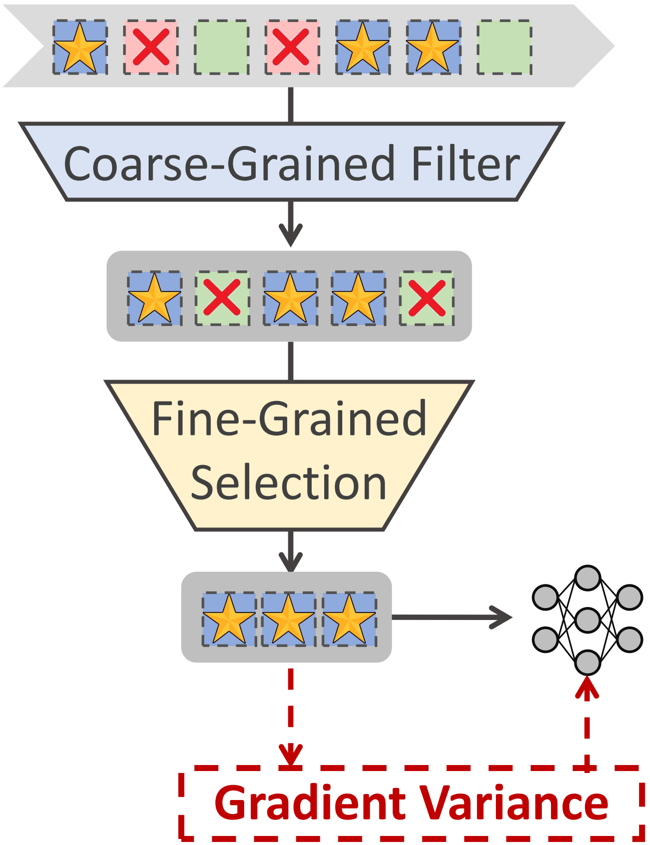
***ms*-level evaluation latency per sample!**

- **Representativeness:** $\mathrm{Rep}(x,y) = -\left\| f_w(x) - \mathbb{E}_{\mathcal{P}(x'|y)}\left[f_w(x')\right]\right\|_2^2,$
  high closeness to class centroid ⟶ **preserve class-level property**.

- **Diversity:** $\mathrm{Div}(x,y) = \mathbb{E}_{\mathcal{P}(x'|y)}\left[\left\| f_w(x) - f_w(x')\right\|_2^2\right]$
  difference to other data ⟶ **reflect sample-level distribution**.

Fine-Grained Selection

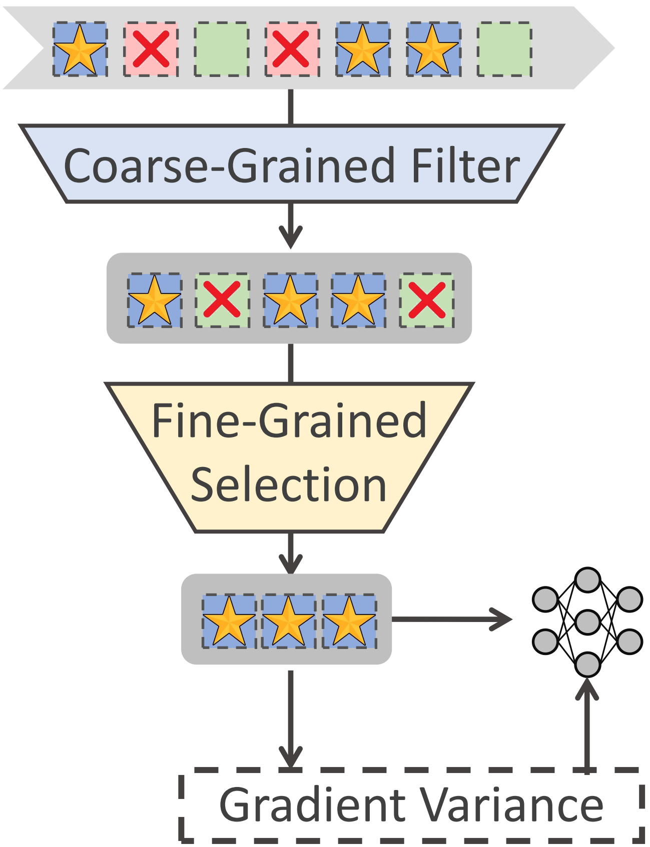
Select data batch with the highest performance improvement.



- **Theorem 1**: Training performance is **inversely correlated** with gradient variance of the selected data.

Fine-Grained Selection (Cont.)

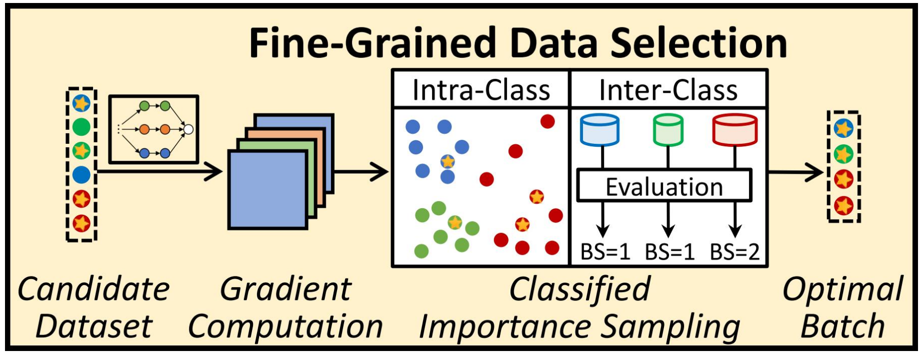
Select data batch with the highest performance improvement.



- **Theorem 2**: Gradient variance of data batch is decided by **inter-class size** allocation and **intra-class sample** selection → **optimal selection policy**.

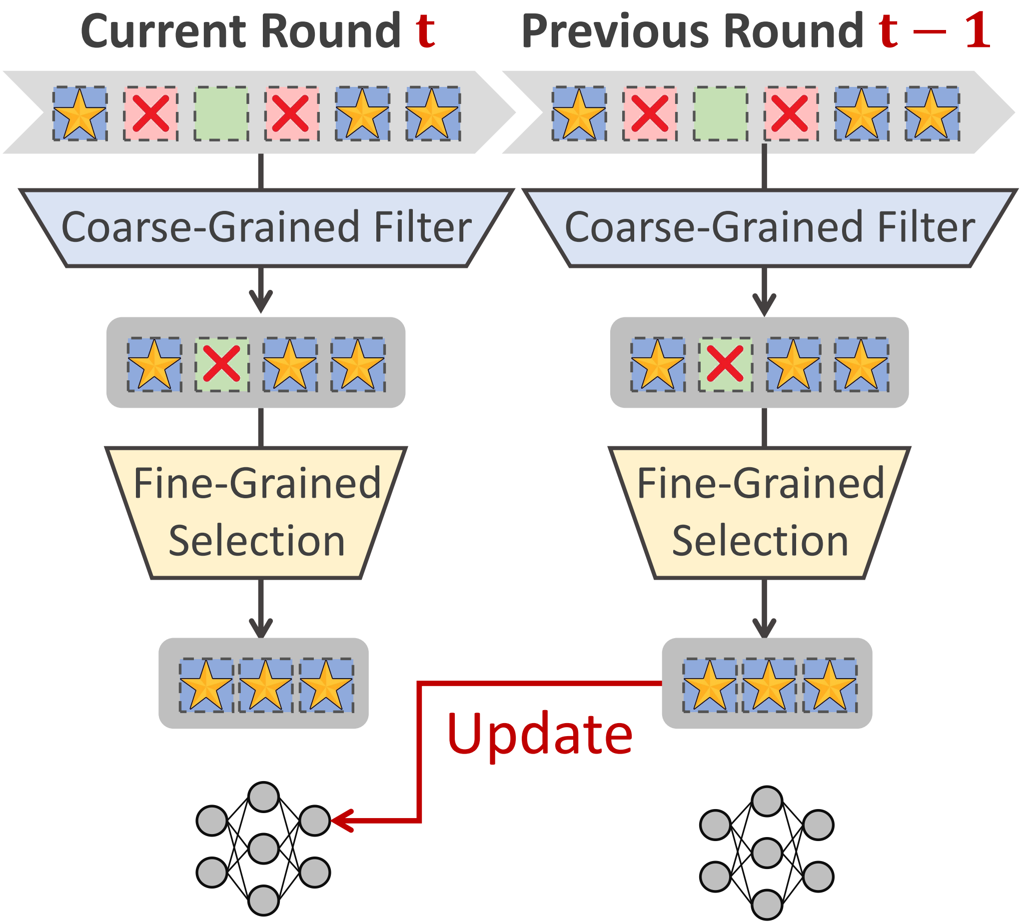
Class Importance: $|\mathcal{S}_y|\Big[\underbrace{\mathbb{V}_{(x,y)\sim P_{t,y}}[\nabla l(w_t,x,y)]}_{\text{variance of gradient}}-\underbrace{\mathbb{V}_{(x,y)\sim P_{t,y}}[||\nabla l(w_t,x,y)||_2]}_{\text{variance of gradient norm}}\Big]^{\frac{1}{2}}$

Sample Importance: $\underbrace{\left\|\nabla l(w_t,x,y)\right\|_2}_{\text{gradient norm}}.$

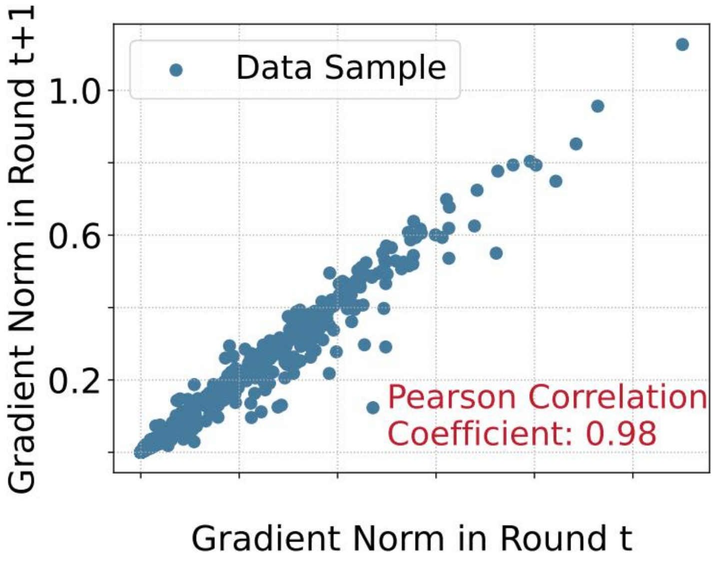


Pipeline Design

Execute data selection and model training in parallel.



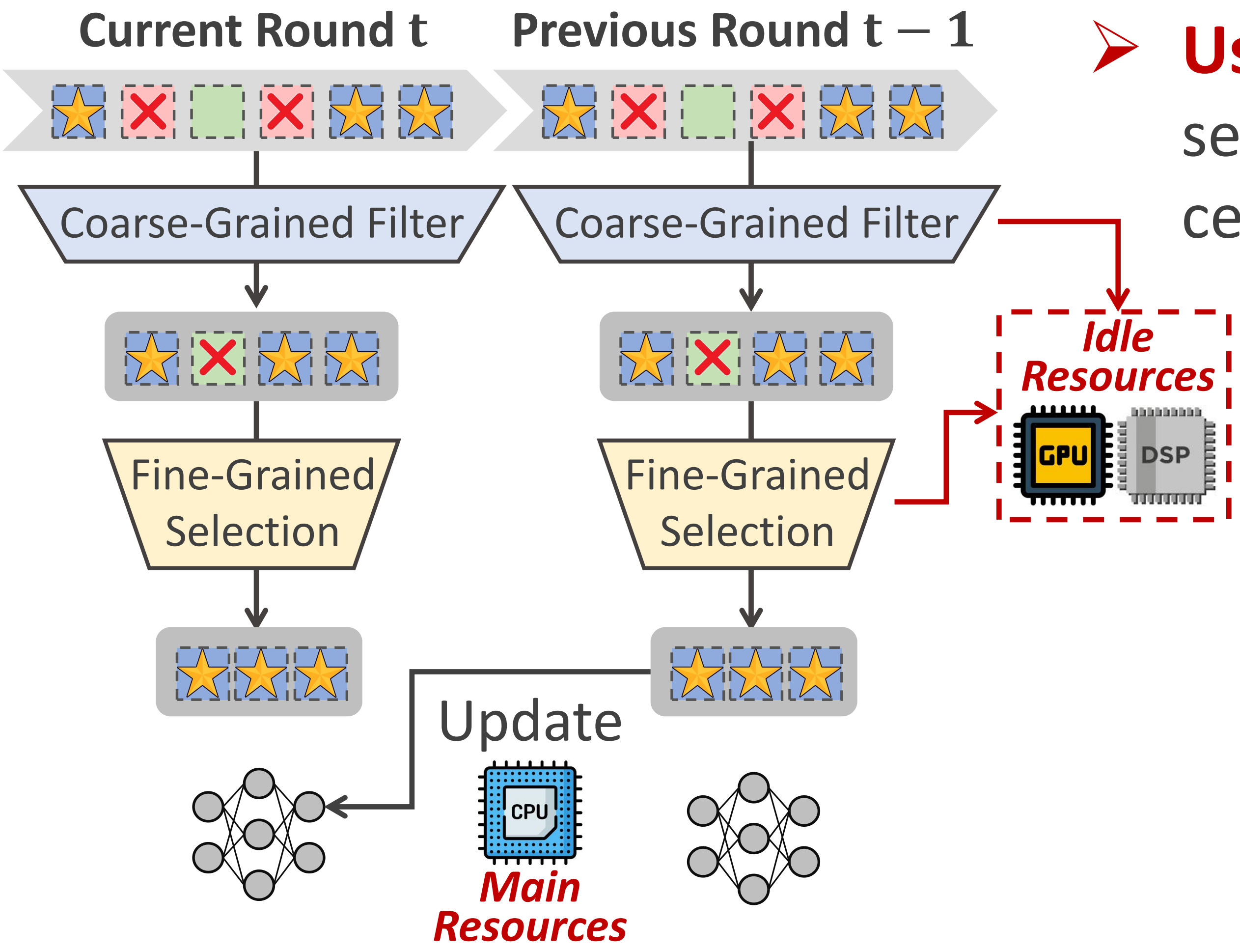
- **One-Round-Delay**: Model in round t is updated with data from last round, while selecting data for next round.



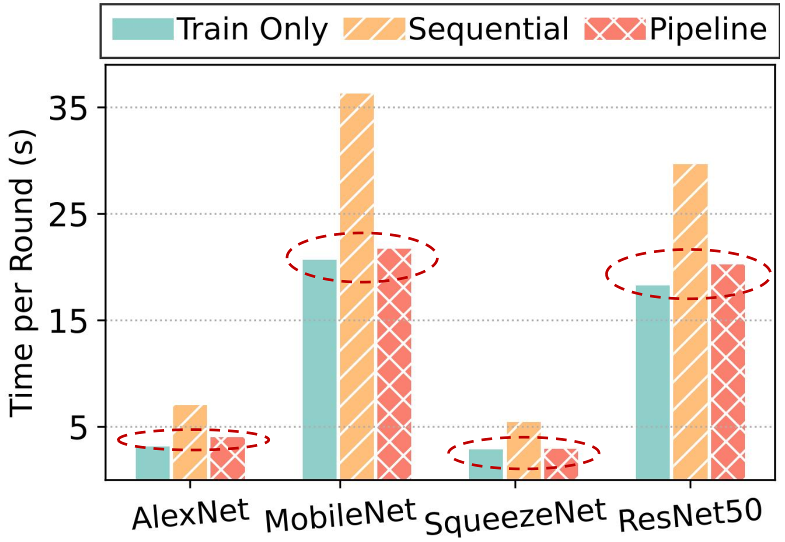
Stable per-sample gradient

Pipeline Design (Cont.)

Execute data selection and model training in parallel.



- **Using Idle Resources**: offload data selection to idle computing resour-ces (e.g., GPU, DSP, NPU)



90% data selection cost is masked

Outline

1. Background
2. Challenges
3. Titan Design
4. **Evaluation**

Setup

3 Tasks & Datasets

- **Image Classification**:
  - CIFAR-10 (60k images of 10 classes)
  - AlexNet, MobileNetV1, SqueezeNet, ResNet50
- **Audio Recognition**:
  - Google Speech Command (100k sound files of 20 commands)
  - ResNet35
- **Human Activity Recognition**:
  - HARBOX (34k IMU samples)
  - MLP

6 Baselines

- Random Sampling (RS)
- Importance Sampling (IS)
- Heuristic Selection: loss, entropy, representativeness&diversity
- Coreset Selection: Camel

Device Implementation

- **Device:** Jetson Nano (4GB RAM, 4 CPU cores, a Maxwell GPU)
- **On-device Data:** 100 samples/round to simulate high speed setting

Model Training Performance

Reduce wall-clock training time to reach target accuracy.

Task	Model	Normalized Time-to-Accuracy (×)							
		RS	IS	LL	HL	CE	OCS	Camel	Titan
IC	AlexNet	1.00	3.25	3.98	3.98	3.59	4.06	2.07	0.70
	MobileNet	1.00	3.22	3.45	3.45	3.41	3.67	1.15	0.57
	SqueezeNet	1.00	3.96	3.97	3.97	3.04	4.06	2.07	0.69
	ResNet50	1.00	2.32	3.14	3.14	2.20	2.18	1.11	0.66
AR	ResNet34	1.00	2.04	3.14	3.14	2.96	3.19	0.81	0.77
HAR	MLP	1.00	3.56	6.30	6.47	5.28	14.4	12.5	0.71

Training Speedup

- Image Task: 30%-43%
- Audio Task: 23%
- HAR Task: 29%

Maintain or improve final accuracy of on-device model.

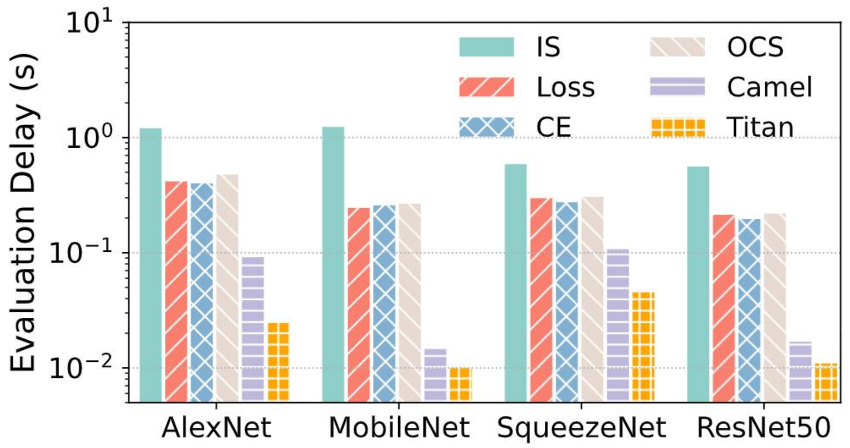
Task	Model	Final Model Accuracy (%)							
		RS	IS	LL	HL	CE	OCS	Camel	Titan
IC	AlexNet	71.2	73.5	18.2	34.3	71.6	62.3	71.3	74.5
	MobileNet	69.2	69.5	17.7	13.9	69.6	38.1	68.7	75.4
	SqueezeNet	76.2	73.0	18.3	45.0	78.0	40.7	75.6	79.0
	ResNet50	76.5	78.0	22.3	34.9	81.7	27.3	76.8	81.1
AR	ResNet34	76.0	78.7	14.7	58.8	73.2	59.4	76.5	79.8
HAR	MLP	75.5	77.5	45.5	21.8	60.9	68.0	75.6	76.7

Accuracy Improvement

- Top 1 accuracy for most models
- Top 2 accuracy for other models

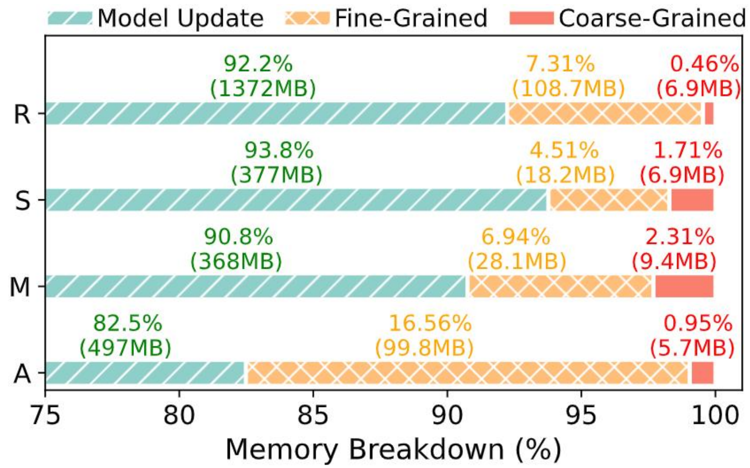
System Overhead

Latency (ms)



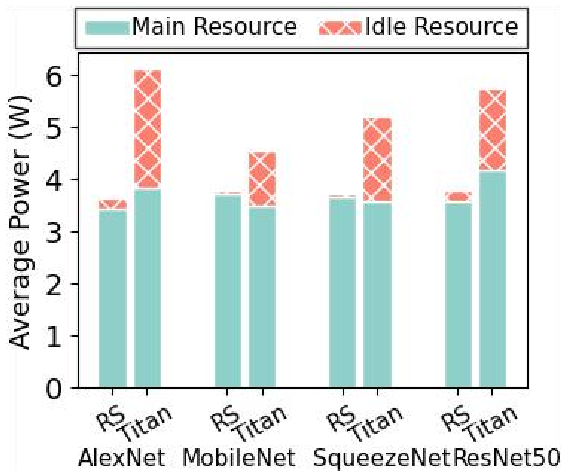
Processing streaming data with **4-13 ms latency**

Memory (MB)



Marginal extra **memory footprint (≤120MB)**

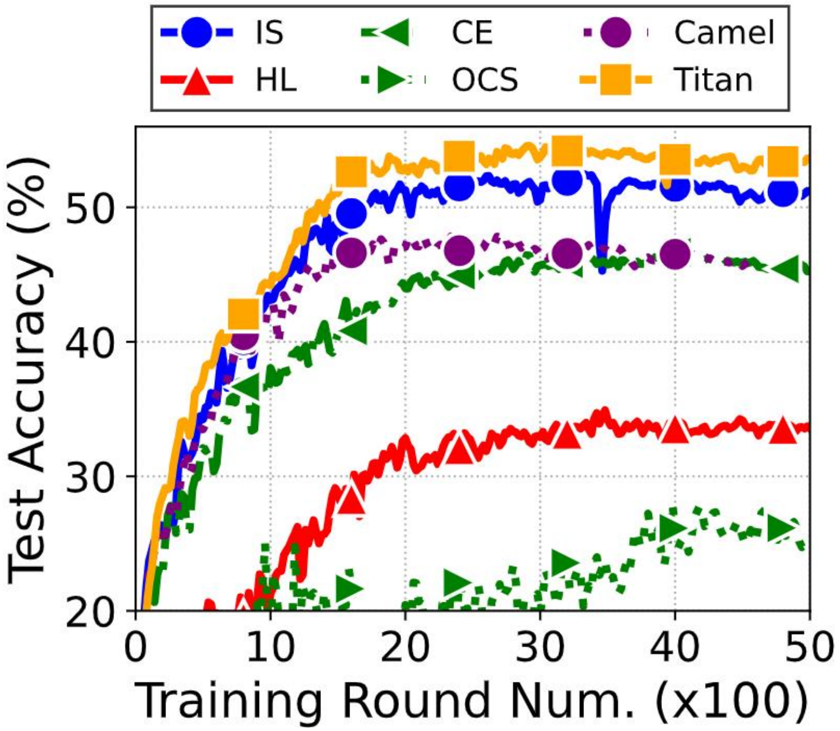
Power&Energy



Higher device power but **lower overall energy**

Extended Scenario: Federated Learning

Improve convergence rate and final accuracy of global model.



Settings: Train MobileNetV1 on cifar10 dataset non-uniformly distributed on 50 devices.

Results:

- **2.03% increase** in global model accuracy
- **3.17x speedup** in number of rounds to reach convergence

Conclusion

Problem

- The **data utilization bottleneck** in on-device model training.
- Existing solutions show ineffectiveness and inefficiency.

Solution

- Titan, a two-stage data selection framework that simultaneously achieves **efficiency and effectiveness.**

Result

- Titan shows **superior training performance** in different tasks with varied data modalities with **marginal system overheads.**

Conclusion

Problem

- The **data utilization bottleneck** in on-device model training.
- Existing solutions show ineffectiveness and inefficiency.

Solution

- Titan, a two-stage data selection framework that simultaneously achieves **efficiency and effectiveness.**

Result

- Titan shows **superior training performance** in different tasks with varied data modalities with **marginal system overheads.**

Thank You for Your Attention !

Chen Gong
gongchen@sjtu.edu.cn